TEC-0051

US Army Corps of Engineers
Topographic Engineering Center

ARPA Unmanned Ground Vehicle Stereo Vision Program

H. Keith Nishihara
Stanley J. Rosenschein
Matthew Turk
J. Brian Burns
Hans Thomas
Monnett Soldo

Teleos Research
576 Middlefield Road
Palo Alto, CA 94301

March 1994

Prepared for:
Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209-2308

Monitored by:
U.S. Army Corps of Engineers
Topographic Engineering Center
7701 Telegraph Road
Alexandria, Virginia 22310-3864

T E C

U.S. ARMY TOPOGRAPHIC ENGINEERING CENTER

Destroy this report when no longer needed.
Do not return it to the originator.

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

The citation in this report of trade names of commercially available products does not constitute official endorsement or approval of the use of such products.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1994	Final Report Dec. 1991 - Dec. 1993

4. TITLE AND SUBTITLE

ARPA Unmanned Ground Vehicle Stereo Vision Program

5. FUNDING NUMBERS

DACA76-92 C-0005

6. AUTHOR(S)

H. Keith Nishihara J. Brian Burns Stanley J. Rosenschein
Hans Thomas Matthew Turk Monnett Soldo

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Teleos Research
576 Middlefield Road
Palo Alto, CA 94301

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Advanced Research Projects Agency
1400 Wilson Boulevard, Arlington, VA 22209-2308

U.S. Army Topographic Engineering Center
7701 Telegraph Road, Alexandria, VA 22310-3864

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

TEC-0052

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This report documents work carried out at Teleos Research in support of the UGV Demo II mission. This research has been divided between basic studies relating to binocular stereo, shape recognition and technology development. Teleos' approach to the study of visual perception embodies a strategy for focusing on the minimal form of perceptual measurement that is meaningful or useful. Several new stereo techniques were developed under this research program including: a way to automatically set stereo matcher operating parameters, an analysis of stereo disparity gradients, and a method for improving area correlator performance in the presence of disparity gradients. Several new techniques for shape detection and recognition are reported. An active vision system for tracking moving objects applies results from the stereo and shape recognition work. A number of efforts were directed at supporting the UGV mission, including the use of narrow-field-of-view stereo for vehicle navigation; very high resolution wide-field-of-view stereo to support landmark based navigation; test data collection for comparison and performance evaluation; and collaborative efforts with the other UGV stereo contractors at SRI and JPL to foster technology transfer.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Shape recognition, active vision, UGV	69
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

Contents

Preface

This research is sponsored by the Advanced Research Projects Agency (ARPA), 1400 Wilson Boulevard, Arlington, Virginia 22209-2308 and monitored by the U.S. Army Topographic Engineering Center (TEC), Alexandria, Virginia 22310-3864, under Contract DACA76-92-C-0005, by Teleos Research, Palo Alto, CA 94310. The Contracting Officer's Representative at TEC was Ms. Linda Graff. The ARPA point of contact is Charles Shoemaker, Program Manager for the Demo II Program.

1 Introduction

This report reviews the work done at Teleos Research during December 1991 through December 1993 in support of ARPA's Unmanned Ground Vehicle (UGV) program. Teleos' research activities have been divided between basic studies relating to binocular stereo, shape recognition, and technology development relevant to the UGV mission. Highlighting the results of this two year research effort are a number of new techniques for enhancing stereo matcher performance on UGV relevant imagery and some new algorithms for detecting and describing shape in range or intensity imagery.

Section two discusses Teleos' approach to the study of visual perception. A strategy for focusing on the minimal form of perceptual measurement that is meaningful or useful is presented. It is argued that this approach is effective for developing efficient, practical systems for carrying out perception tasks. This view is then related to Teleos' work with stereo matching and accelerator design.

Section three introduces several new stereo techniques developed under this research program including: (1) techniques for automatically setting stereo matcher operating parameters such as filter size by previewing results on a sparse set of points over a range of possible parameter settings; (2) identifying the principal parameters affecting the magnitude of the disparity gradient effect that compromises correlator performance in UGV stereo imagery; and (3) improving area correlator performance in the presence of large stereo disparity gradients.

Section four describes new techniques for shape detection and recognition developed under this program. They include: (1) using stereo to rapidly discriminate figure from background during real-time tracking; (2) representing shape structures based on local concentrations of image mass; and (3) rapidly detecting and classifying simple grouped structures in two-dimensional imagery.

Section five describes work done on active vision where the stero and shape discrimination work were applied in a real-time system for tracking moving objects. In this work, high bandwidth servo control systems were developed to use frame rate stereo and motion information to drive an active stereo camera head. This tracker module was then installed on a mobile robot cart. The cart was then controlled by another servo system guided by range and heading information from the tracker module.

During this period Teleos carried out a broad range of activities oriented toward applying results from the core research in stereo and shape recognition to the needs of the unmanned ground vehicle program. Section six of this report reviews this work which includes: investigation of the use of narrow-field-of-view stereo in support of vehicle navigation; use of wide field-of-view, high-resolution stereo mosaics to support navigation using stereo landmarks; test data collection for comparison and evaluation of stereo algorithms; and collaboration with the UGV stereo contractors at SRI and JPL and technical liasons at the UGV system integrator, Martin Marietta to transfer stereo sensing technology to the UGV program.

2 Task directed visual perception

One of the goals of our research has been to develop and understand practical demand-driven computer vision. Our view is that the sophisticated performance observed in biological systems is, to a large degree, derived from the fluent use of simple and robust measurement capabilities. We are attempting to identify and study modular perceptual abilities in biological systems that fit this model. We have worked extensively with two such modules: stereo disparity measurement and optical flow measurement.

Research on stereo and motion sensing and enabling processing technology has matured to the point where interesting real-time applications such as vehicle guidance and security and surveillance are practical. Three broad questions are pertinent to the design of visual perception capabilities supporting applications like these: (1) what specifically should we measure, (2) what are the best algorithms to use, and (3) what is the best hardware technology to build on?

In the following subsection, we present an approach to thinking about perceptual measurement which makes the claim that simpler is better. The idea here being that a few good measurements directed by the needs of a particular task can be as effective (or more so) as a much larger number of undirected measurements. This idea casts visual perception as an activity involving the application of a small set of measurement tools.

Subsections 2.2 and 2.3 then describe briefly how our work with stereo and motion processing relates to this minimal-meaningful-measurement paradigm.

2.1 Minimal meaningful measurement tools

The perceptual information a blind person needs about his environment and the character of aids that prove most useful to him can provide practical guides for research in machine vision. There is a close analogy between the sensing needs of an intelligent blind person and those of active problem solving machines.

To be acceptable for use by a blind person, a visual aid must be easy to use, informative, and cost-effective. Interestingly, aids that are "too smart" are often rejected because they leave the blind user oblivious to much of the detail of what is going on. This makes it hard for that person to use the tool effectively in new contexts. What seems to be called for, in the case of the

blind, are aids that such users can operate as tools to accomplish perceptual tasks.

Following this line of thought, a desirable perceptual aid for machine vision ought to recover some basic information and it should have an easy-to-model behavior that is sufficiently rich to allow an expert to use it in creative ways. A blind person's cane is a good example: it has a consistent mechanical behavior and it provides timely information about the presence or absence of physical objects at dynamically selected locations about the operator. The cane "device" has low-bandwidth input and output interfaces to the user—that is, manual pointing control and force, vibration, and sound feedback. This allows it to be managed easily by the blind user while carrying on other parallel activities such as conversation. Furthermore, though simple, the cane has a fairly rich and consistent behavior that fosters the development of expertise in its use. For example, one learns the *feel* of different pavement textures or conditions—slippery or uneven.

We think of a measurement tool as a device that a higher level agent can deploy "skillfully" in specific task domains, much as a blind person uses a cane or as an artist uses a brush. Three qualities are noteworthy of such tools:[1]

1. **simple but meaningful.** The device should make the simplest meaningful measurement possible to be efficient. Too much automatic interpretation at this level can be counterproductive and too many gratuitous measurements can waste processing resources. This orientation makes it is easier to present more precise information to the user and it allows the user to interpret the basic measurements with increased efficiency and precision.

2. **easy-to-model.** The device should have a consistent, easy-to-model behavior. If the underlying algorithm has many special case behaviors, it becomes difficult for a user to anticipate that device's behavior in new situations or possibly even in familiar ones.

3. **informative output.** The device should exhibit a behavioral richness that encourages the learning of strategies for making more specialized measurements with it. For example, simply reporting best estimates of range from a stereo correlation tool would deprive the user of valuable information about the shape of the correlation peak. In various

circumstances, that user might be able to use knowledge of the peak's height, its broadness in vertical disparity, or its bimodality.

This measurement tool concept can be applied to the study of early vision problems to help us define computational problems that are somewhat different from the problems that are traditionally addressed. For example, instead of attempting to compute a dense stereo range map, we concentrate on the problem of computing and communicating the results of a single range measurement over a patch of surface. This distinction can be significant when issues of interaction with higher level knowledge and control are considered.

In stereo matching, a measurement over a small sensing area may fail due to the absence of matchable features. To recover, the calling agent can try switching to a larger measurement window or it can move the original measurement patch to a slightly different position, or it might decide to move the sensor head to a better vantage position. In either case, the calling agent is aware of the changes made and their implications for the measurement. It is in possession of knowledge of the task to be accomplished, it is aware of the measurement difficulty and the character of the possibly degraded information obtained. At the same time this agent does not have to know much about the detailed workings of the measurement algorithm itself. As long as it exhibits a consistent and predictable behavior it can be used effectively when treated as a black box.

2.2 Sign-correlation image matching theory

The first class of computations studied extensively following the mimimal-meaningful-measurement approach have been image matching algorithms applicable to stereo range finding and optical flow field measurement. A computational theory for measuring stereo and motion disparity was developed that is consistent with the measurement tool objectives. In this subsection, the sign-correlation theory is briefly described. The following section will then explain how it is implemented in a demand directed system following the methodology of section 2.1.

Binocular stereo, the measurement of optical flow, and many alignment tasks involve the measurement of local translation disparities between images. Marr and Poggio's zero-crossing theory made an important contribution towards solving this disparity measurement problem[2]. The zero-crossing

theory, however, does not perform well in the presence of moderately large noise levels as has been illustrated by the inability of zero-crossing based approaches to solve transparent random-dot stereograms—which, interestingly, can be perceived correctly by the human visual system[3]. A sign-correlation algorithm that builds on Marr and Poggio's ideas and that addresses many of the weaknesses of the original work has since been developed.[4]

We continue to use the zero-crossing primitive for matching, but the matching rule is changed. Instead of matching zero contours, we correlate the signal's *sign* in an area. This subtle change makes a significant difference in the behavior of the matcher. Sign-correlation continues to provide useful disparity measurements in high noise situations long after the zero-crossing boundaries, surrounding the signed regions, cease to have any similarity. An intuitive explanation of why the two approaches perform so differently follows from the fact that the sign of the convolution signal is preserved near its peaks and valleys long after increasing noise has caused the zero contours to be fully scrambled. Thus, area correlation of the sign representation yields significant correlation peaks even with signal-to-noise ratios of 1 to 1. Since sign-correlation still operates off of the zero crossing representation, the key strengths of Marr and Poggio's theory are preserved.

2.3 Acceleration for real-time operation

The sign-correlation algorithm has been implemented in a demand-directed hardware architecture following the approach described in section 2.1. Teleos' current implementation known as Prism-3 incorporates a video rate $\nabla^2 G$ convolver and a fast area correlator that can make a set of 36 parallel correlation measurements at different disparities at any computer designated image location.

We have made significant advances in developing algorithmic and hardware techniques for accelerating the large kernel convolutions and area correlations used by the sign-correlation approach. At present a pair of VME bus boards along with a general purpose processor board carry out full frame stereo convolutions at video rate with $\nabla^2 G$ convolution operators as large as 60 by 60 pixels. Stereo disparity measurements covering a search space of 72 by 4 pixels at a typical resolution of .2 pixels are accomplished in 300 microseconds.

Area-based motion measurement can be done using the same facility with

an interframe search range of 108 by 108 pixels and the same subpixel resolution, in under 2 milliseconds. This allows the tracking of bodies moving at angular rates of 30 degrees per second with a sensor having a 10 degree field of view. That translates to being able to detect and follow a subject entering the field of view traveling at 36 km/h at a range of 10 meters.

Optical flow fields and stereo range measured sparsely over the entire visual field can be used to do rapid figure ground discrimination. This result can then be used to focus attention and further processing on meaningful physical entities.

3 Stereo core research

During the course of the research effort we completed four tasks aimed at improving the performance of a stereo matching system such as our sign-correlation system when operating on UGV-relevant imagery. These efforts resulted in (1) an automated technique for selecting pre-processing filter sizes appropriate for dynamically varying scene characteristics; (2) an analysis of a disparity gradient effect which can have a significant effect on stereo matcher performance in UGV imaging configurations; (3) a skewed correlation window technique for efficiently mitigating the disparity gradient effect; and (4) an evaluation of our matching algorithms carried out in cooperation with other UGV stereo team members.

3.1 Control parameter selection

The principal control parameter of the sign-correlation approach is the size of the convolution filter used. This parameter selects the spatial scale at which texture is picked up for use in the stereo correlation. In many cases a very large operator accentuates texture that is more stable than that available at finer scales. Since the correlation window size scales with the filter size, it is desirable to find the smallest filter size that yields acceptable correlation measurements.

To automate the selection of an appropriate operator size we developed an algorithm that pre-samples the stereo image at a small number of locations and at each of these it checks the quality of the stereo correlation peak over a range of filter sizes.

Specifically the algorithm does a search over filter size, w, from the set: {4, 6, 8, 12, 16} (units are pixels). For each filter size, correlation statistics are sampled at 25 locations evenly spaced over the filtered stereo pair.

At each of the 25 sample locations, we search for the highest correlation peak in the disparity search range. We also keep track of the second highest peak. For each of these two peaks we compute the peak's *local height* above the correlation values at a distance in horizontal disparity of $.75w$ from the peak disparity. This prevents blank regions in the images from looking good. We then take the difference between the local height of the best peak and the second best peak at the sample location. This *peak-difference* allows us to fold in a requirement that there not be multiple peaks at the sample location.

The smallest filter satisfying at least two of the following three criteria was selected:

1. Average peak-difference score is above 0.3.

2. At least 90 percent of samples have peak-differences better than 0.1.

3. At least 80 percent of samples have peak-differences better than 0.25.

If no filter size satisfies this test, the filter with the largest average peak-difference is used.

This algorithm exhibits the following characteristics:

1. It favors smaller filter sizes which yield better spatial resolution. If all else is equal, we do better with smaller filter sizes which allow us to use correspondingly smaller correlation window sizes.

2. It enlarges filter size when shot noise levels are high. Shot noise is prevalent in low contrast areas of an image and in many night vision sensors. When these levels get sufficiently large relative to the local image texture contrast, matching performance drops. A larger filter often will increase the relative strength of coarse texture patterns relative to this type of noise.

3. It enlarges filter size to avoid ambiguous correlation peaks. In some imagery there is repeating structure, such as checkerboard patterns, that cause ambiguous correlation peaks. These repeating patterns are sometimes not present at coarser scales because they are dominated by other coarse scale variations, such as slight irregularities in the checkerboard.

This filter size preselection technique operated effectively on the large JISCT evaluation project discussed in a later section of this report.

Figure 1 shows a stereo pair prepared by CMU to illustrate the ambiguity problem that binocular stereo matchers can run into. The shoe appears to be sitting on a rubber door mat but is actually held several inches above the mat. The repetitive pattern present in the mat allows most matchers to incorrectly follow the false correlation surface that continues from the shoe's sole out over the door mat. CMU researchers have argued that solving this type of problem requires the use of multiple baseline stereo.

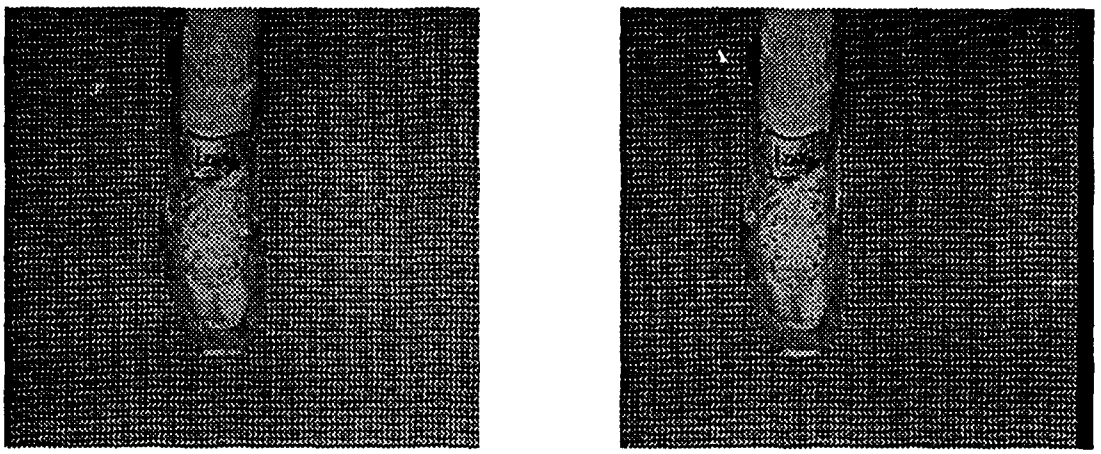

Figure 1: CMU Shoe stereogram illustrating ambiguity problem for binocular stereo.

Figure 2 shows the sign representation obtained from figure 1 when filtered with a medium sized $\nabla^2 G$ operator. As one would expect, the repeating pattern on the door mat shows up crisply and the sign correlation surface on the mat exhibits multiple ambiguous peaks. Interestingly, when our filter size selection algorithm was applied to the CMU shoe stereogram, it selected a much larger filter size which yielded the sign representation shown in figure 3.

With this larger filter the repeating pattern of the door mat has been replaced by a coarser texture pattern associated with the pattern of irregularities in the mat. Thus unambiguous disparity measurements can still be made using binocular images with the larger filter size. Figure 4 shows a disparity surface plot computed using the sign correlation algorithm. There is some rounding of the surface at the shoe edges due to the large correlation windows used. This averaging effect can be reduced by following the coarse matching step with subsequent passes using smaller filter sizes using the coarse data to disambiguate the multiple correlation peaks.

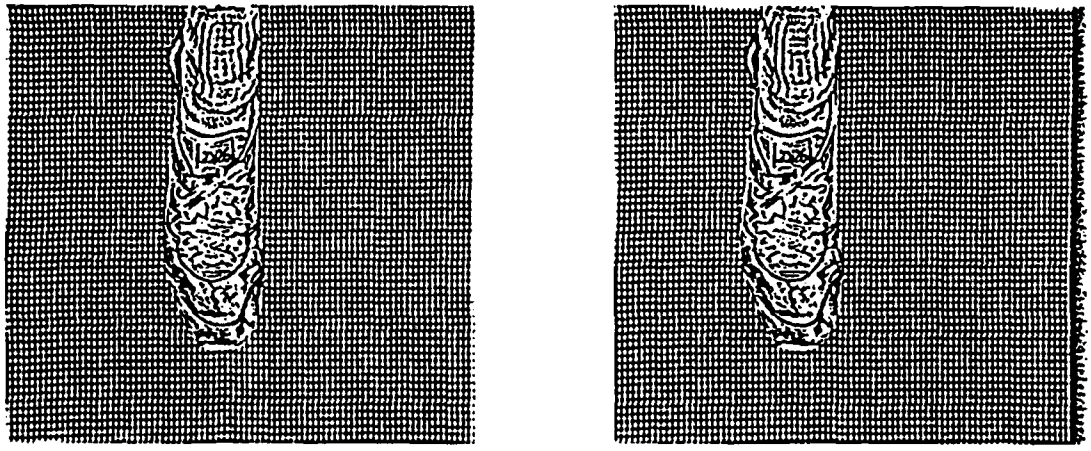

Figure 2: Sign representation of figure 1 using a medium sized filter.

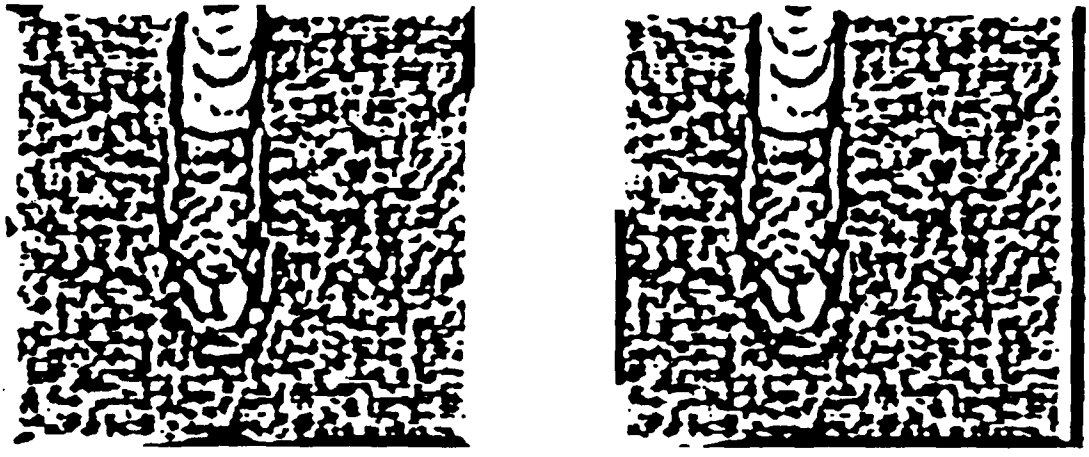

Figure 3: Sign representation of figure 1 using a filter 3 times larger than that used in figure 2.

Figure 4: Disparity surface plot of the CMU shoe stereogram computed using the sign representation shown in figure 3.

3.2 The disparity gradient effect

A new and rather surprising result regarding stereo disparity gradients in the UGV imaging configuration was obtained. These gradients occur when the cameras view an inclined surface such as the flat road out in front of the vehicle as depicted by figures 5 and 6. They can significantly affect the performance of area correlation based matchers because the receding surface under a correlation window does not register at any single disparity. This causes the correlation peak obtained to be lower and spread out making detection of the peak more difficult and unstable.

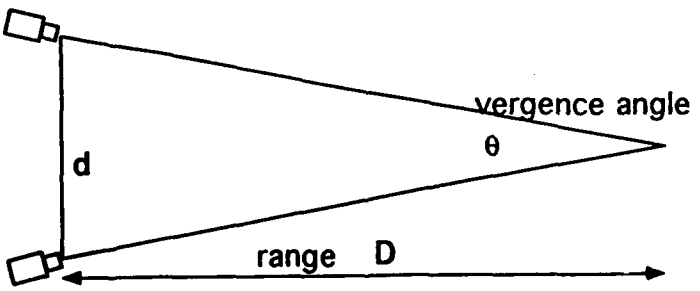

Figure 5: Aerial view of stereo imaging configuration.

An estimate for the disparity gradient magnitude was derived as a function of stereo sense head parameters such as camera lens size, baseline separa-

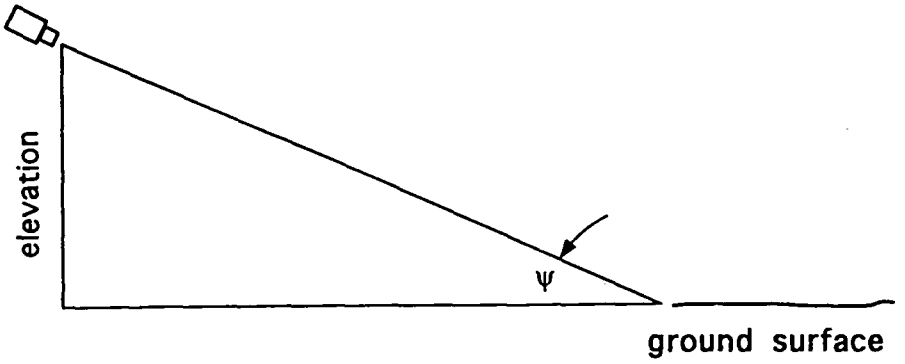

Figure 6: Side view of stereo imaging configuration.

tion between the cameras, camera height, and the pitch angle of the cameras (how much down from the horizon are we looking). The expression obtained is:

$$disparity\ gradient \approx \frac{baseline}{height} \quad (1)$$

In other words, the disparity gradient depends primarily on the ratio of camera separation to camera height. It does not depend significantly on lens size, or pitch angle of the cameras so long as the cameras are looking significantly farther ahead than they are high above the ground.

An important consequence of this result for the UGV program is the constraint it imposes on the baseline separation. Typical matching algorithms are seriously affected by gradients larger than about 0.2 pixels disparity change per pixel in the image. This means that camera separation should not be larger than one fifth of the height of the camera head above the ground.

We derive the approximation of equation 1 using the notation on figure 7 as follows:

$$\frac{elevation}{D} \approx \frac{\Delta s}{\Delta D} \tag{2}$$

$$\approx \frac{\Delta \alpha D}{\Delta D} \tag{3}$$

Equation 2 is by similar triangles, assuming that Δs is much smaller than D and that Ψ is small.

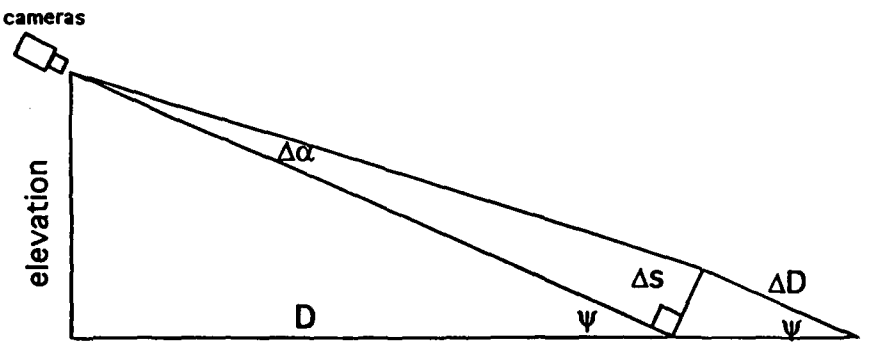

Figure 7: Relation of vertical disparity to vertical image position over a flat inclined surface at a distance.

Rearranging terms allows us to write:

$$\frac{\Delta D}{D} \approx \frac{\Delta \alpha D}{elevation} \tag{4}$$

Now from other work [4] we have the relation:

$$\frac{\Delta \theta}{\theta} \approx \frac{\Delta D}{D} \tag{5}$$

Combining with equation 4 we get:

$$\frac{\Delta \theta}{\theta} \approx \frac{\Delta \alpha D}{elevation} \tag{6}$$

rearranging and substituting $\frac{baseline}{D}$ for θ we get:

$$\frac{\Delta\theta}{\Delta\alpha} \approx \frac{\theta D}{elevation} \tag{7}$$

$$\approx \frac{baseline}{elevation} \tag{8}$$

which is the desired relation.

This information has been communicated to the UGV system integrator and these considerations will be reflected in the sensor head design. Teleos is continuing to work on methods for extending the operating envelope of its matching algorithm to handle larger disparity gradients.

3.3 Skewed correlation window technique

Compensation for vertical disparity gradients can be made when a correlation measurement is made by progressively shifting the horizontal disparity between the left and right correlation windows as we scan vertically over those windows as shown in figure 8. Adjusting the correlation window "skew" can greatly improve the correlation peak height obtained in images with large vertical disparity gradients. The window skew that gives the best correlation can also be used to directly estimate the local disparity gradient. A similar operation can be done to compensate for horizontal disparity gradients. A future firmware upgrade to the Prism-3 hardware accelerator will allow dynamic compensation for both vertical and horizontal disparity gradients.

Figure 12, in the following section, shows an example of a stereo disparity map computed on a surface with a large vertical disparity gradient. The skewed correlation window technique was used in this case and it yielded almost a doubling of correlation peak height.

3.4 Performance evaluation

A critical component of the stereo research and development program for the UGV mission is the definition of evaluation metrics and tests. As one of Teleos' first projects on the program, it collaborated with the other UGV stereo contractors to formulate test protocols and collect test data. The team assembled a large test database of stereo imagery, known as the JISCT

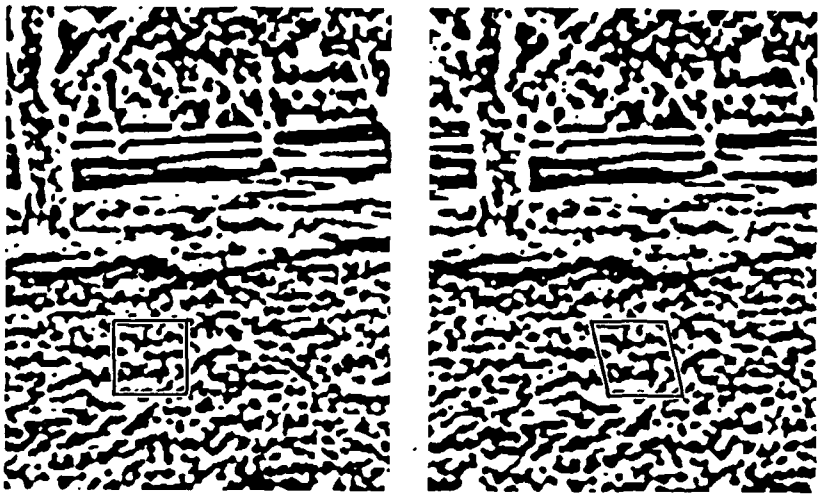

Figure 8: A skewed correlation window is shown on the right that compensates for the vertical disparity gradient between the two images.

database, and ran several of our best matching algorithms on them for a comparative analysis. The test stereo pairs were distributed to participating groups in early January. Results using algorithms operating at SRI, Teleos, and INRIA (in France) were collected for analysis.

In support of this evaluation effort, Teleos developed a set of stereo test images designed to evaluate performance on scenes with large vertical disparity gradients and high noise levels. Both factors are typical of UGV imagery. This suite of test images are all of the same scene, a flat steeply inclined surface (68 degree incline) with a spherical object at the center of view. The first stereo pair of the set, shown in figure 9, is with good exposure settings on both cameras. Figure 10 superimposes graphs of the intensity profile along a horizontal raster line through the tennis ball in the left and right images of figure 9. The ball is slightly shaded and gives rise to the abrupt dip in both curves near the center of the plot. With scrutiny one can observe the correlation between intensity fluctuations in the two superimposed graphs.

Figure 11 shows the Laplacian-of-Gaussian sign patterns computed from the stereo pair in figure 9. The filter size—automatically selected—had a center diameter of 6 pixels. Figure 12 plots the disparity surface computed from correlating the sign patterns. The cameras were to the left and the line of sight is horizontal in the display. Note the flatness of the surface around

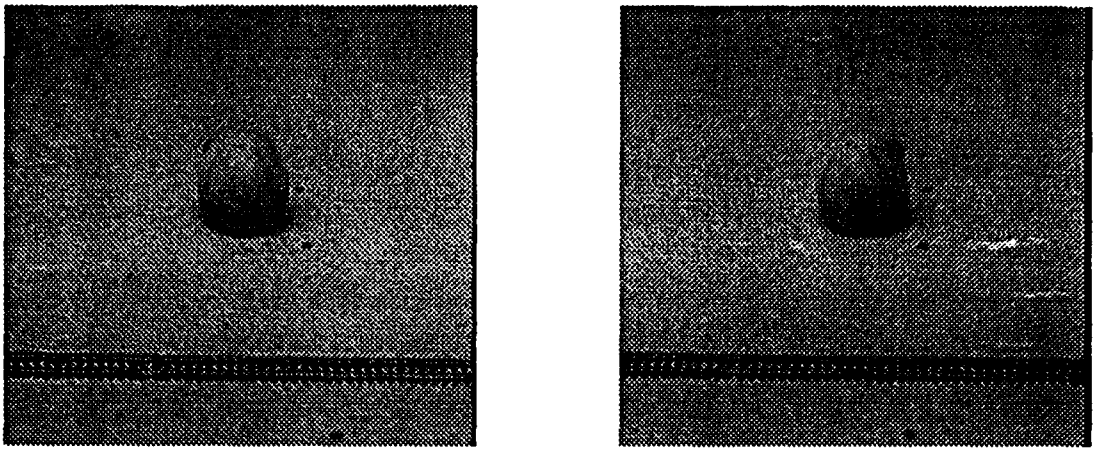

Figure 9: Stereo pair of a flat board at 68 degree incline to camera axis with a tennis ball at the center.

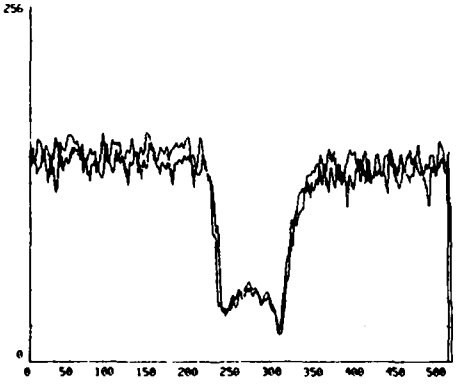

Figure 10: This graph superimposes plots of the camera intensity along a raster line passing through the center of the tennis ball from both images in figure 9.

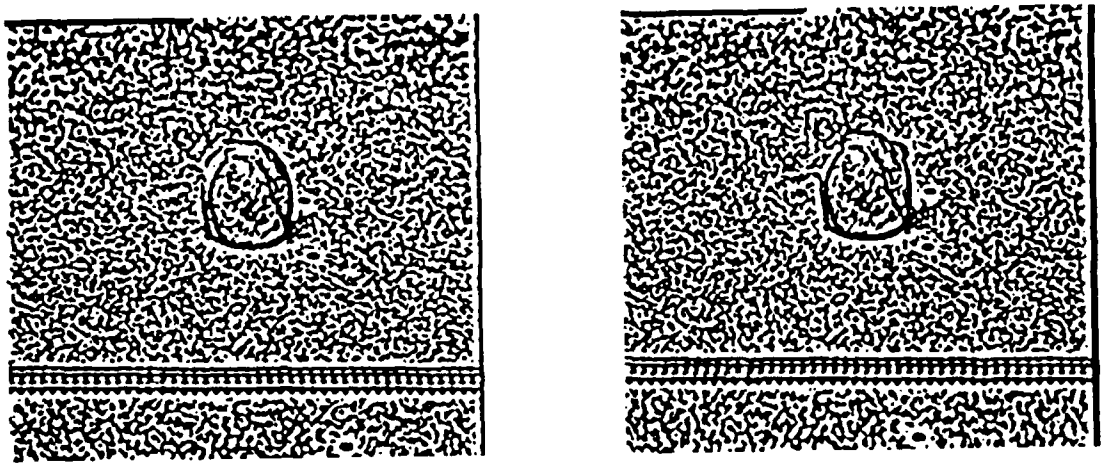

Figure 11: Sign representation of figure 9

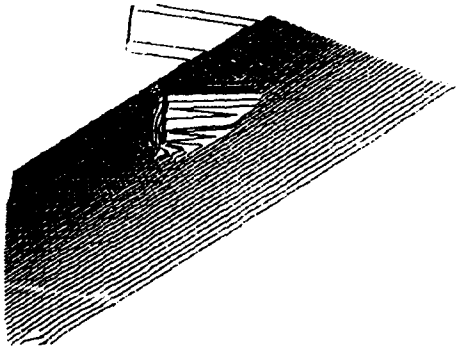

Figure 12: Disparity plot computed using the sign correlation algorithm from the sign representation shown in figure 11.

the ball at the center. Unmatched regions—due to unsatisfactory correlation peak height or shape—were mapped to zero disparity mostly around the perimeter of the ball and in a small patch at the back edge. Ninety seven percent of the image was matched.

Each successive pair in the test suite of seven stereo pairs has the camera aperture reduced by one stop on the lens aperture. This created a series of stereo images with the same stereo scene but increasing noise levels due to the decreasing amount of light available to the cameras. The last pair in the set has no useful signal present. As we expected, most stereo algorithms were able to handle the first pair of this set, but performance of individual algorithms fell off at different points in the series of test pairs.

Figures 13 through 16 show the same sequence of displays for the fifth pair in the test suite. Figure 13 shows the raw stereo pair, but here the lenses have been closed down 4 stops and we see an essentially black display. There is still a small amount of contrast remaining, as can be seen from the graphs in figure 14, but the ratio of signal to sensor and digitization noise is decreased by a factor of 16. Figure 15 shows the sign patterns obtained from the raw stereo images. In this case a slightly larger filter center diameter of 8 pixels was automatically selected for use in the correlator. Note that moderately stable patterns are discernible even at this increased noise level. Figure 16 plots the disparity array computed in this case. Ninety percent of the attempted points yielded acceptable correlation peaks and as can be seen from the plot most of the board surface was recovered. Similarly, the ball's curvature is still apparent. Increased dropouts occurred at the edges of the image and around the ball's occluding contour. By contrast, no other algorithm participating in the JISCT evaluation was able to get any meaningful results on this stereo pair.

For the next image pair after the one shown in figure 13, with 50 percent less light, the sign-correlation algorithm still yielded matches at half of the image locations.

As noted earlier, the sign-correlation algorithm was the only one to perform well on the repeating background of CMU's *shoe* stereograms. As we saw in figures 1 through 4 these stereo pairs were intended to be an example of a matching problem that could only be solved by multiple baseline approaches. We found, however, that the coarse-to-fine scale space filtering available with our sign-correlation approach allows the problem to be solved with a single stereo pair with any baseline. Sign-correlation succeeds

Figure 13: This stereo pair, the fifth in the test suite, was taken with the lens' closed down 4 stops relative to figure 9 and it is essentially black.

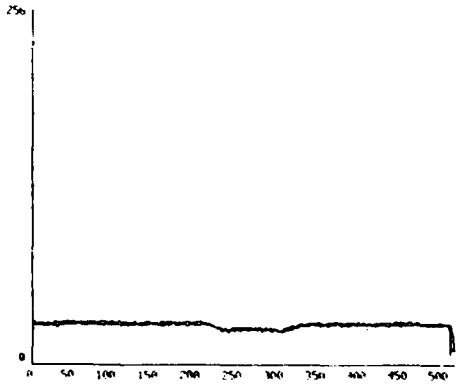

Figure 14: This graph plots the camera intensity for figure 13 at the same position as in figure 10.

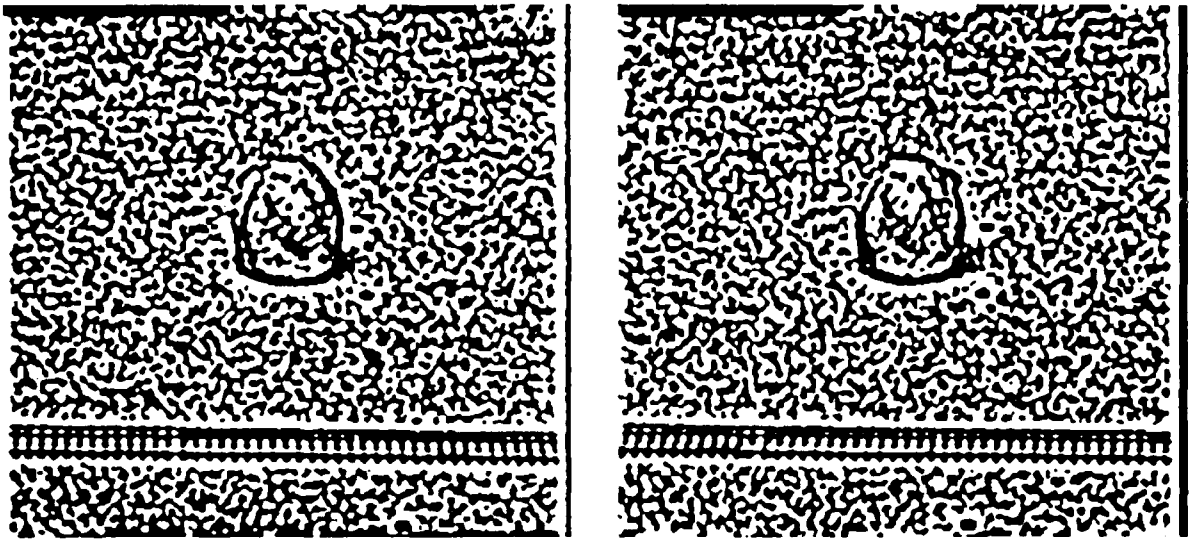

Figure 15: The sign representation of figure 13 still shows moderately well correlated structure despite the sixteenfold loss of contrast.

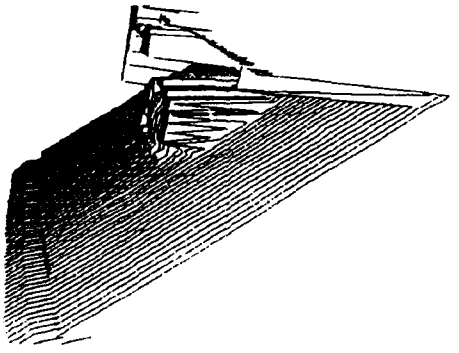

Figure 16: The disparity plot obtained from figure 15 is essentially the same as that shown in figure 12.

in this case because the wallpaper illusion in the example is not perfect, there are small variations due to dirt or other irregularities in the floor mat that was used for the background of the scene and that introduces coarse scale textures that can be used by our matcher to make unambiguous disparity measurements.

A comparative analysis of the test results by SRI indicated that Teleos' sign-correlation approach performs very well on ground terrain and in large noise situations, and as expected, slightly less well than some other approaches at discerning small objects due to the larger filter and window sizes employed.

4 Shape recognition core research

An important component of Teleos' research effort has been to study ways to extract information about objects either positive or negative—for example, holes—from earlier sensory processes such as stereo, motion, or color. This section first presents an overview of our approach to the study of shape representation and recognition. The second subsection describes a blob description algorithm that addresses the question of efficiently detecting and representing geometrical structure in an image such as a range map. The third subsection, on a scale-space ridge representation, examines one aspect of the blob description problem in greater detail.

4.1 Shape representation and recognition

Much research has been done in the field of shape recognition though very little has yet been accomplished with respect to the fluent recognition of common three-dimensional shapes in everyday settings. Marr and Nishihara[5] started work on a theory that had at its heart a simple shape description module. That work shared many of the characteristics of the present work. The intent was to produce a hierarchical primitive shape description incorporating up to seven positional elements at each node of the hierarchy. For example, at the scale of the whole body, a human shape might be characterized by the positions of the major appendages (arms, legs, head). This level of detail is too limited to allow identification of a particular individual, but it is sufficient to name a shape category and to identify shape components that could be recursively described by the same primitive shape module.

The Marr-Nishihara paper formulated the issues and outlined an idealized representation. There remained major unsolved problems, however, with computing the desired shape primitives and with determining the proper grouping of those shape primitives in the construction of their hierarchical shape representation. There have been sufficient advances in early-vision processing over the past decade to allow a continuation of that line of research productively.

The research is aimed at establishing the hypothesis that fluent shape recognition can be achieved with rapidly-computed minimalist descriptions in conjunction with a hierarchically structured database of known shape categories.

4.1.1 Place-marker primitives

A very simple description of shape indicates just the center position of a body's mass along with an estimate of its average diameter. While this amount of information would not suffice to discriminate very many objects, a slight variant of the idea has much more potential. Consider locations in the shape of an object where a set amount of its mass is locally concentrated. A sphere is uniformly shaped and no local concentrations of mass can be identified. However, a football shape has two pointed ends where its mass is locally pinched out, forming "bumps." If a way could be found to define robustly and then actually identify these mass concentration centers reliably, the geometrical arrangement of those positions could be used to help classify the shape. For example a football is distinct from a sphere in that it has an extra pair of mass-centers.

Continuing this chain of thought, how is a watermelon shape different from that of a football? Both exhibit a pair of mass concentrations ("bumps") to either side of the body center. The football, however, continues to exhibit strong mass centers at very fine scales since its ends are pointed while the melon ceases to have any appreciable local mass concentrations at fine scales since its ends are rounded. This idea is reminiscent of Blum's[6] symmetric axis representation, but it differs in several respects. First, we are interested in averages (or best fits) for the mass centers rather than positions where the spheres fit exactly within the shape's envelope; second, we are interested in obtaining a small number of place-markers—the local mass centers—one scale at a time.

A central question addressed by our research has been to discover an effective way to compute mass-centers from visual data. Following the ideas of Crowley and Parker[7], our preliminary experiments on two-dimensional silhouettes looking at peaks in Laplacian of Gaussian convolutions have been encouraging.

4.1.2 Hierarchical matching

Once we have constructed a hierarchical representation of a shape in terms of the positions of its "bump" place markers at various scales, how can we efficiently index into a large database of objects to find the most similar object? The hierarchical organization of "bumps" can speed this process

considerably. Matching can begin at coarse scales, at which each object only has a small number of bumps and hence can be matched extremely quickly. Database objects which match the input shape well at coarse scales can then be matched to the shape at finer scales, while objects that do not match well at coarse scales can be removed from consideration.

4.1.3 Figure-ground from stereo and motion

Sensory modalities such as stereo, motion, color and texture can play a key role in defining areas within an image that go together as a unit. This figure-ground separation has the potential in shape recognition applications to simplify greatly the combinatorial grouping problems that inevitably arise in cluttered imagery.

Figure-ground separation tasks do not require very high spatial resolution, and hence are well matched to the capabilities of existing robust real-time stereo and motion algorithms developed by Teleos. It is interesting to note that whereas conventional stereo or motion based approaches have strived to produce high-spatial-resolution data, human psychophysical performance while being very robust, appears to have fairly low spatial resolution in stereo and motion. This supports the notion that human stereo and motion mechanisms are principally attentional aids for figure-ground separation, and that approaches concentrating on computing high-resolution data may be poorly matched to these measurement modalities.

The figure-ground discrimination potential of our real-time stereo and motion modules presents a promising opportunity to simplify the grouping problem in shape recognition significantly. A major component of our work has been directed at using those measurement modules to assist with grouping and figure-ground discrimination.

4.1.4 Robust low-resolution measurements

Range measurements made with the sign-correlation algorithm are area averages over the correlation windows we use. Since the user knows this, he can take advantage of how such average measurements behave as the correlation window is swept over different physical shapes. Knowledge of the tool's behavior coupled with the user's knowledge of his visual world should allow him to request measurement sequences that answer domain specific

questions, such as where is the table edge with high precision[8].

By addressing a simple but useful demand driven task, we find that we can make better measurements, and that we can more easily integrate the use of such measurement tools into application development without compromising the theoretical integrity of the underlying methodology. In several instances, the computational theories we have developed following this approach have held up well as biological models for human vision[3, 9, 10].

In light of this, our goal for shape recognition has been to strive to find the simplest modular measurements that obtain principled results. We think of these modules as measurement tools, that are capable of providing single primitive perceptual measurements such as stereo range or optical flow and bump-based shape characterizations in a time frame matched to the rate at which a user—or application program—can reasonably be expected to ask questions during the coarse of accomplishing a visual task.

4.1.5 Classification by place-primitive geometry

In summary, the working hypothesis is that: (1) the generic and sufficiently expressive position markers can be computed from images; (2) the arrangement of a small number of such markers can be used to discriminate classes of shapes; and (3) a hierarchical application of this place marker technique can be used to rapidly refine shape classifications in large databases. The distinguishing characteristics of this approach are use of coarse but robust descriptions at multiple scales, and application of a recursive description/lookup process over a viewed shape and its sub-parts to acquire sufficient information to accomplish a given recognition task.

The following sections describe progress in the areas of building structural descriptions based on blob representations and extracting stable position markers using a scale-space ridge concept.

4.2 A blob description algorithm

This section describes a blob representation and detection scheme that looks at the broad question of bringing bottom-up sensory information together with higher-level object model-based constraints.

The blob-detection scheme uses a pre-compiled array of elliptical blob masks with fixed positions, sizes, and orientations. The scheme is intended

to be fast and is designed to work with relatively sparse measurement data. A blob family consists of a set of these blob masks at a given scale at regularly spaced positions in the image. Several families may be active at the same time. Figure 17 shows a sub-family at a given scale and position—it includes three stretches (or elongations) and four orientations. The crosses of Figure 18 shows the image locations marking the centers of each blob sub-family.

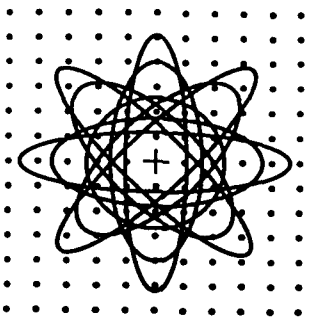

Figure 17: Family of blob masks with different orientations and elongations centered about one point.

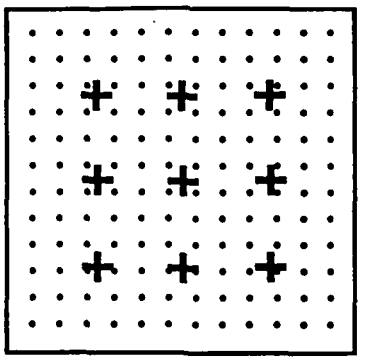

Figure 18: Image locations where blob families shown in figure 17 are replicated.

The salience of each blob is determined by some function of a relatively small number of image measurements which fall within the blob's boundary (and optionally within the boundary of a related surrounding blob, the blob's "surround"). These measurements are centered at fixed locations, such as the

dots in Figure 18. Each measurement location may fall within the extent of several blobs, as in Figure 19. Because these locations and the blob locations are fixed for a given blob family, we can pre-compile a list of blobs affected by each measurement location. Then, when running the algorithm, each measurement value directly updates the appropriate blob descriptions.

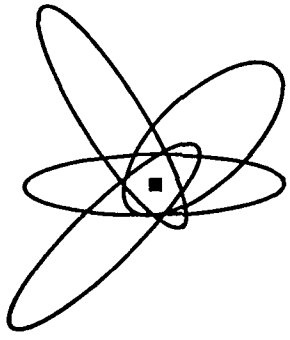

Figure 19: A given image pixel can fall into the *receptive field* of several blob masks as illustrated here.

The blob description accumulates the sum and sum of squares of the image measurement values, so that after one pass of the image it can quickly calculate the mean and variance for each blob. The salience or distinctness of each blob is then calculated as a function of the mean and variance of the measurement data. The blobs are then ranked according to their distinctness values, with the most distinct blobs chosen to initially represent the object(s) of interest.

There are several design issues in the algorithm, such as what scales are most appropriate, how many blob locations to allow per family, how many measurement locations to choose for a given blob scale, how many stretches and orientations to use, and how to define the distinctness function. Also, the algorithm as stated does not use information from past frames—it is run anew each time. How can information from the previous frame or frames be used to increase speed of blob extraction or tracking, and to increase reliability? How can the blobs be aggregated to form more meaningful, parts-oriented descriptions? And how can multiple sources of measurement data—e.g. color, intensity, range, motion—be integrated or coordinated to take advantage of low-level vision measurements provided by subsystems such as

the Prism-3 system?

The following sections describe in detail the issues, experimental results, and progress in the development of a blob detection mechanism to support recognition.

4.2.1 Sampling

The measurement data may consist of range data, motion vectors (not yet implemented), color, and grayscale pixel data. These measurements are computed from different sized neighborhoods in the original input image(s)—from a single pixel to a large window surrounding the output "location." In early experiments with grayscale image data (using dark objects on a bright background) the blobs were quite unstable. This was due to undersampling. A single pixel value was being sampled for each region which led to a large amount of aliasing. This type of instability was not seen when using range data as the measurement input, since the depth information is calculated using large convolution windows.

To improve the performance using direct image data (color and grayscale), the measurement layer returns an averaged area, rather than a single pixel value, essentially low-pass filtering the image data. The area size is selectable, but defaults to a value that fits the selected sampling sparseness. This slows down the measurements slightly, but significantly improves the blob stability.

4.2.2 Spacing

There is a tradeoff between the number of fixed blob locations (and therefore accuracy of the representation) and the computational expense of computing the salient blobs. The coarser the description (i.e. the larger the blob size), the fewer blob locations are necessary, and conversely, the finer the description, the more blob locations are necessary. For example, if a blob is 100 pixels wide, another blob located just 8 pixels to the right has significant overlap, and the measurement data used to calculate the blob distinctness may actually be equivalent (since the measurement data is also discrete and sparse). However for blobs that are 12 pixels wide, a shift of 8 pixels significantly changes the underlying extent of the blob.

The blob spacing is selectable when a blob family is created. The default is a multiple of the blob diameter, so that both large and small blob families

are spaced properly.

Spacing is also an issue for tracking blobs, since a blob identified in one frame should be found nearby in the next frame. (See Section 4.2.4.)

4.2.3 Distinctness measures

The first measure of blob distinctness studied was simply average value (intensity, color, depth, ...) inside the blob. This is useful for simple, well-modeled scenes such as silhouettes or objects protruding in depth, but is not applicable to most scenes. More useful is a "center-minus-surround" measure which assigns a high distinctness value to blobs whose center covers an area of one average measurement value (brightness, color, depth, ...) and whose surround covers an area of a significantly different measurement value. This distinctness measure may use any combination of measurement averages and variances to produce a distinctness estimate.

Figure 20 depicts an example of blobs at two scales. The underlying data is a high-contrast grayscale image of a toy person. The best (highest distinctness values) small blobs are found along the arms and legs where their scale fits best, and the next highest distinctness values are found in blobs alongside the torso (not shown). The best larger blobs cover the torso (and sometimes the head).

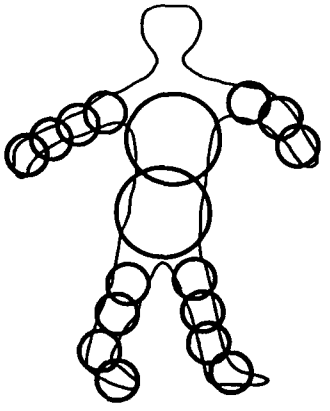

Figure 20: Illustration of a circular blob description of a human figure using two blob scales.

4.2.4 Tracking

The blob families are originally created to cover the whole image, with spacing intended to describe the object shape at various scales. The algorithm to calculate salient blobs is data-driven—no information from previous frames is used.

To track moving objects, a faster, more precise locating method is needed to update previously found blobs. The same technique can be used with minor changes. Rather than using a blob family with relatively sparse locations across the whole image, the family is altered so that blob locations are densely located in a small area near the previous blob location. Figure 21 describes this situation —in Figure 21(a) the initial blob is located at one of the cross locations, sparsely distributed throughout the image. But in the tracking mode, as in Figure 21(b), the blob locations are more densely distributed around the previous location.

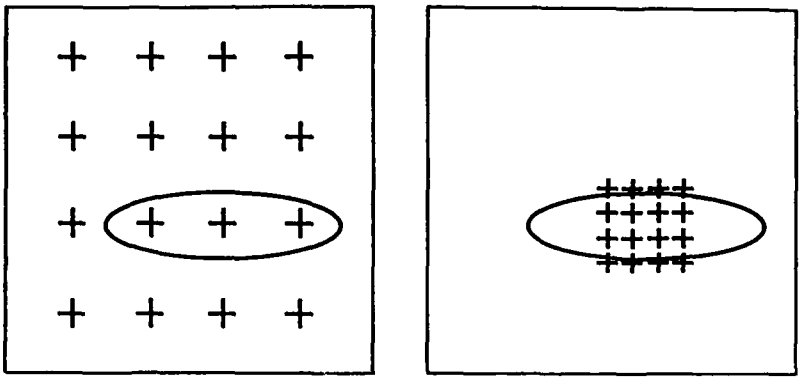

Figure 21: When tracking a blob found in the frame shown on the left, blob detection resources can be allocated more densely about the position it was last seen in as shown on the right.

4.2.5 Blob grouping

To create meaningful representations from an initial set of salient blobs, the blobs must be aggregated to form intermediate-level part descriptions. In the case of a human figure, for example, we would like blob-driven representations of the arms, legs, torso, and head. The problem is to transform an initial grouping of blobs covering a part into a single representation of that

part. Minimal length description methods for efficiently representing parts are being studied for this purpose.

4.2.6 Multiple measurement sources

Certain measurement sources are intuitively well-suited for particular imaging situations: e.g., grayscale image data for high-contrast scenes, color data for finding uniformly colored objects, range data for protruding objects or objects separated in space from the background, motion vectors for moving objects, etc. In more general viewing conditions, however, objects should be modeled and recognized based on their salient properties, whether known in advance or not. Ways are currently being investigated for combining, rather than choosing among, measurement sources. This should significantly increase the generality and robustness of these techniques.

4.2.7 Using circles rather than complete blob families

Initial experiments with blob detection based on the complete blob families indicated that the elongated (stretched) blobs were not very stable. The distinctness values of several blob orientations and stretches were typically very close, and a small change in the input image caused a different choice of the most salient blobs. On the other hand, circular blobs were much more stable. The restriction to circular blobs should increase robustness with little computational penalty since a blob aggregation step is already a part of the blob description process.

4.2.8 Analysis

Our experiments with this fairly general prototype architecture has led us to the conclusion that our early intuitions (see section 4.1) concerning shape representation seem appropriate: (1) blob distinctness is well seen as a significant difference between internal and external image values; and (2) general shape is can be well seen as organization in a collection of localized place-marker primitives (which rely, for example, on a single circular blob mask shape).

4.3 A scale-space ridge representation

The results above motivated a more careful analysis of the problem of detecting local concentrations of *stuff*. This is treated as a two step process of first applying a circularly symmetric detector and then applying a technique for organizing the output of the symmetric detector to recover a more global description of shapes in the underlying image.

The representation presented in this section builds on the notion that figures of interest tend to be regions with significant contrast in some feature value relative to the surrounding image. The initial development of these ideas has been carried out on simple intensity images and the following presentation uses examples from that domain. These concepts have natural extensions to other sensing modalities such as stereo, motion, and color imaging.

A description that in some way makes explicit the way image mass is clumped in an image—for example, indicating that the image has a central bright region—can be useful for discriminating figures from background and as a means for decomposing complex figures into simpler components. Two important criteria for such a representation are that it allows scale specific description of structure and that it provide information that has good stability characteristics.

4.3.1 Approach

When an image is smoothed—for example, with a large Gaussian convolution—fine detail is suppressed leaving a blobby image that highlights areas that are bright or dark at the scale of the Gaussian filter. Ridges in such smoothed images correspond to the centers of elongated regions that are light relative to a background, and valleys similarly correspond to relatively dark elongated regions.

The size of image structures that remain in the Gaussian smoothed image is related to the Gaussian's space constant, σ. Following the Gaussian smoothing operation with a high-pass filter creates a band-pass filter which accentuates structures with a mean diameter of about 2σ. This can be appreciated by considering the result of convolving a 1D image of a white box on a black background with the second spatial derivative of the Gaussian (G_{xx}). This operator has a negative center zone and a positive surround.

When the negative/positive zone boundaries match the white/black region boundaries, the operator response is of maximum magnitude. This occurs when the region width is exactly twice σ.

Hence, image mass centers can be localized in the image by (1) detecting ridges (or analogously valleys) in the Gaussian filtered image at different scales, and (2) determining if each ridge is a true image mass center via analysis of the G_{xx} filtered image at that point. More specifically, a ridge point detected at scale σ is considered the center of an underlying image mass of width 2σ if the magnitude of the G_{xx} filtered image at that point and scale is a local maxima with respect to scale. In other words, at the ridge point, the G_{xx} output magnitude should grow smaller as the σ varies from the scale at which the image ridge point was detected.

For the above to work properly, the G_{xx} convolution must be normalized by the integral of the magnitude of the convolution operator, $\|G_{xx}\|$ to allow cross scale comparisons. This value varies by $1/\sigma^2$. Therefore normalization can be accomplished by multiplying the operator output by σ^2.

An analogous method of detecting ridge points in two-dimensional images is described and demonstrated below.

4.3.2 A 2-D ridge representation

A 2-D ridge point is defined as the result of the following five step process: (1) convolve the image with a range of Gaussians, (2) detect ridges in each output, (3) select ridges that have a local maximum magnitude second derivative with respect to σ, (4) keep those above a global contrast threshold, and (5) connect neighboring ridge points into ridge spines. For this discussion, the term *ridge* applies to both ridges and valleys.

Figures 22-24 show an image processed in this way. Figure 22 shows the image. Figure 23 shows the selected ridge points (only ridge points associated with light regions relative to their surround are shown). Each ridge point is represented by a circle with a radius equal to the σ at which the point was detected. Figure 24 shows the continuous sections of ridge spines that were extracted. The ridge spines together with the estimated width at each point can be used to represent the shape, position, orientation, and scale of each extracted image part.

The following is a more detailed description of each of the ridge computations:

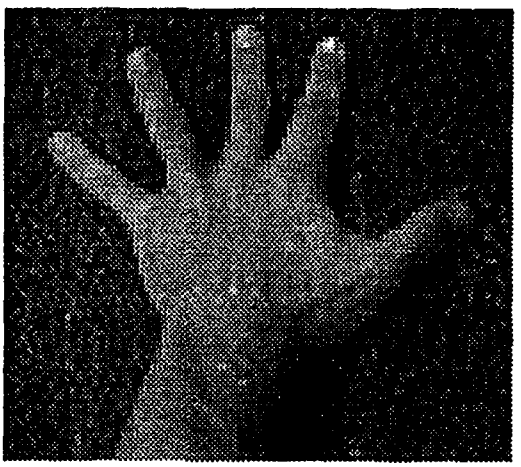

Figure 22: Gray level image of a hand.

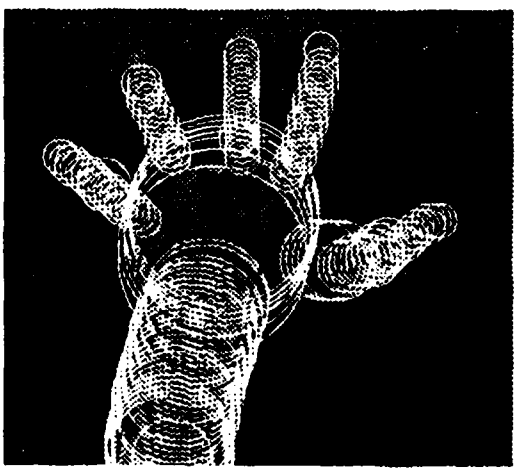

Figure 23: White circles indicate locations of ridge centers on the hand image from figure 22—radius indicates the space-constant, σ, of the operator that gave the strongest response.

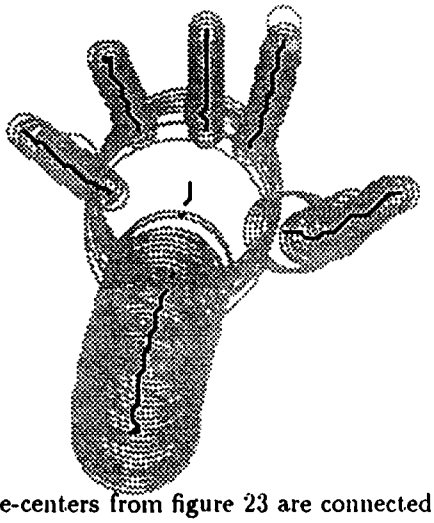

Figure 24: The ridge-centers from figure 23 are connected to show ridge spines.

1. The image is convolved with a set of Gaussians with different space constants (σ). The σ's increase in size exponentially, the smallest and largest are chosen to reflect the range of scales of interest. For the demonstration shown in Figures 22-24, the minimum σ is 1.95 pixels, the maximum is 25.27 pixels, and there are five intermediate values: 3.0, 4.59, 7.0, 10.77, and 16.5.

2. For each σ, ridge points are detected. The first and second derivatives are estimated at each point in the filtered image by fitting a 2D quadratic to the 3x3 neighborhood about the pixel being tested. Solving the eigensystem of the Hessian matrix, the directions of maximum and minimum second directional derivatives, and the values of these derivatives, are found. If the first derivative, in the direction of the maximum magnitude second derivative, is zero at a point sufficiently close to the pixel in question, the pixel is considered a ridge point, with a *ridge curvature* equal to the maximum magnitude second derivative. For the current experiment, the distance threshold was set to $\sqrt{2}$. Note that this is not a curvature or gradient magnitude threshold: the surface can be arbitrarily close to planar (as long as it is not planar) with

the same result.

3. Select ridge points that have a local maximum magnitude second derivative with respect to σ. For each detected ridge point, compare the magnitude of the associated ridge curvature with the second derivatives computed at the same point but at the scales immediately larger and smaller than the current one. (Before comparison, all the magnitudes are scaled by the square of the associated σs.) If the ridge curvature is larger in magnitude than the second directional derivatives (in any direction) at the other two scales, select the ridge point as having a scale appropriate for the underlying image region. As can be seen in Figure 23, the selection process produced ridge points at the appropriate scales for most of the object's structure.

4. Select all appropriate-scale ridge points that are above a global contrast threshold. The magnitude of the second derivative times the square of the associated σ is a function of the region contrast. Thus this value is used to compare with the global threshold. For the ridges selected in Figures 23 and 24, the threshold was 17.0.

5. In the final step, contiguous ridge points are linked into ridge spines. As can be seen in Figure 24, almost all of the lengths of all the salient structures are represented as continuous spines. Together with the estimated widths at each ridge point, they represent the salient aspects of the objects part geometry.

5 Active vision core research

Computer vision systems typically exist as a primary input to some higher-level process. Although many systems have been constructed where there is limited or no feedback from the high-level process to the vision system, there is an emerging belief in the vision community that incorporating powerful feedback mechanisms will greatly increase the capability and durability of various vision algorithms. This new area of vision research has been termed active vision.[11, 12, 13, 14, 15]

In this section the Prism-3 accelerator architecture developed at Teleos for carrying out active vision research is described. Then the design of a real-time tracking module using this architecture is discussed. Finally, results from demonstrations of the system tracking human motion from a fixed mount and from a mobile platform and directions for further work are reviewed.

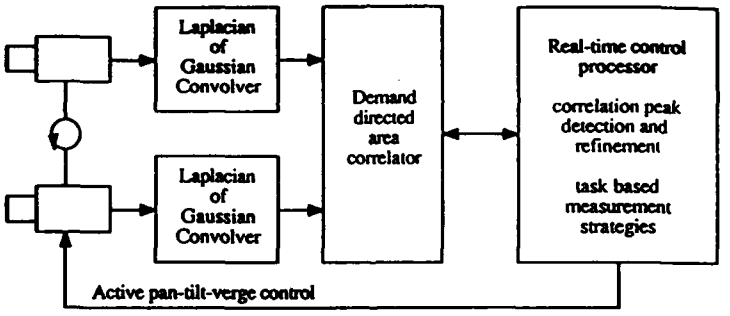

Figure 25: Prism-3 hardware block diagram

5.1 Prism-3 architecture

The sign correlation algorithm has been implemented on the system illustrated in figure 25. The stereo cameras are mounted on an active pan-tilt-vergence mechanism shown in figure 26. The cameras have a stereo baseline of 22.2 cm and the camera vergence angle is computer controlled. The head can move through a 180 degree rotation in under a second and exhibits a positioning repeatability on the order of 50 arc seconds standard deviation in pan, 20 arc seconds in tilt, and 6 arc seconds in vergence.

The two video cameras are operated in a mode where they share the same pixel clock in order to minimize timing skew between the cameras that would result from only using horizontal and vertical video synchronization signals. The left and right camera video is digitized using commercial (DataCube) digitizer hardware. Parallel digital video streams are fed to two dedicated Laplacian-of-Gaussian convolvers (developed by Teleos). These convolvers allow video rate convolution with operator center diameters ranging from 1.6 pixels to 16.6 pixels.

Digital convolution video signals are fed from the two convolvers to a binary correlator board (also developed at Teleos) which carries out high-speed correlations on the sign bits of the input video streams.

The Prism-3 correlator board performs 36 correlations in parallel on rectangular windows of adjustable size. The correlator board is operated by an external control processor (currently a 68040 single board computer). At the start of a measurement cycle, this processor writes the pixel coordinates of the next measurement to be made into registers on the correlator along with information about the disparities at which correlation measurements are to be made. A set of correlations with 32 by 32 pixel windows at 36 different disparities takes 100 microseconds to complete. The correlation results are then read into the control processor. If a well formed peak is identified in the data, quadratic interpolation is used to refine the peak disparity. These steps on the CPU take an additional 200 microseconds.

With correlations taken at even pixel disparities at a single vertical disparity, the above 300 microsecond cycle allows a disparity peak to be located in a 72 pixel disparity search range with a third to a tenth of a pixel resolution. Vertical disparity errors between 1 and 2 pixels are well tolerated.

The correlator hardware is also configured to allow correlations to be computed between successive frames from a single camera. This allows optical flow measurements to be made. In the tracking application described below, the system has been programmed to handle image velocities as large as 50 pixels per frame in any direction with subpixel measurement resolution.

The dedicated hardware incorporates standard off-the-shelf components and makes extensive use of field programmable gate arrays (FPGAs) to achieve high performance while maximizing flexibility in reconfiguring the hardware design.

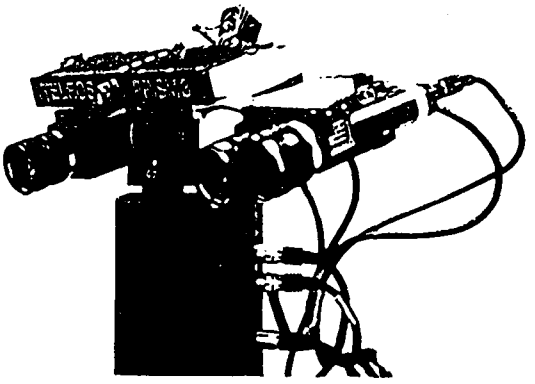

Figure 26: The active head assembly of the Prism-3 system.

5.2 The tracker module

Tracking and control applications require fast low latency response from the sensor to be of value. A natural limit on speed is the frame rate of the camera system; for most commercially available cameras this is either 30 or 60 frames per second.

At 30 Hz, a person three meters from a camera walking across the field of view at 1 meter per second will traverse about 38 arc minutes per frame. With a 50 mm lens the interframe motion disparity will be on the order of 30 pixels. This estimate is for one set of parameters—disparity magnitude varies approximately linearly with lens focal length, subject distance, subject speed, and frame rate—but it gives an indication of the kind of matching performance that will be required to follow human scale motions.

Similarly the head position control must be responsive to velocity commands at that 30 Hz rate with maximum acceleration, and velocity limits sufficiently high to allow smooth pursuit tracking motions.

A tracking system designed to meet these performance specifications was implemented on the Prism-3 architecture as three subsystems, a low-level electronic tracking system, a mechanical servoing system, and a figure sta-

bilization system. These individual mechanisms operate as parallel process threads which are loosely coupled. The electronic tracker makes high performance image based measurements of optical flow and stereo range. The tracker attempts to electronically follow an externally designated patch of surface so long as it remains within the camera field-of-view. The mechanical tracker operates the active camera head in velocity mode using a PID control algorithm. This system attempts to keep the head pointed so that the coordinates of the surface patch tracked by the electronic tracker are kept close to the center of the camera field of view. The figure stabilization submodule uses stereo measurements to assess the extent of the figure associated with the tracked patch. If the tracked patch is not centered on that figure, this module sends an error bias signal to the electronic tracker in an attempt to push it back to the center of the figure. This helps to maintain tracking on figures that are undergoing rotation that would otherwise lead an optical flow based tracking scheme astray.

5.2.1 The electronic tracker submodule

The electronic tracking module measures optical flow velocity and stereo range off of a single window position in the camera image. That window position is initially designated externally. For example, the initial window coordinate might be set to the center of the camera field and started with the cameras pointed at a person to be followed. The Prism-3 correlator is configured to use six of its 36 correlators for stereo correlation and the remainder for measuring optical flow—that is correlations on successive images from the same camera.

A software control routine reapplies the correlator hardware 30 times with different disparity settings to measure correlation against motion at 900 different motion disparities (30 by 30 samples). The samples are spaced apart in disparity space by 4 pixels so this gives a total disparity range of 120 pixels. This means that any motion less than 60 pixels between camera frames will be covered within the search range.

Similarly, stereo correlations are done at 180 disparities (60 by 3 samples). With the four pixel sample spacing in disparity space, this covers a disparity range of 240 pixels horizontally by 12 pixels vertically. Thus as long as the actual disparity at the tracked patch is less than 120 pixels either way it will

be detectable and vertical misalignments as large as 6 pixels can be tolerated.

The above set of 900 plus 180 correlation measurements is made each frame time. If a well formed correlation peak, (i.e. with satisfactory height and sharpness), is located in the optical flow correlation surface, that peak disparity is used to adjust the correlation window position to cause it to follow the measured motion for the next frame's measurements. The current tracking window coordinates are reported to the external process modules.

Likewise if a well formed stereo peak is identified, the disparity measurement, relative to the current stereo fixation plane, is reported.

5.2.2 The mechanical tracker submodule

The head servo mechanism is by its nature a more sluggish system that must cope with inertial mass, power, and vibration constraints that limit its responsiveness. To minimize the influence of these limitations on tracking a moving image patch, the electronic tracker described in the previous section is only responsible for moving a massless window on the image. The mechanical servo system is made responsible for monitoring that window position and servoing as fast as it can to keep that window near the center of the camera field.

The mechanical servo operates in velocity control mode which is natural for this application. For example, if a tracked target is moving to the left across the camera field at 10 degrees per second, the head servo will see the tracked window move off to the left edge of the camera field and it will start to accelerate in that direction. As it does the tracked window will cease to move on the camera image since the head velocity is catching up with its velocity. A PID control algorithm is used to cause the head's turning velocity to exceed the velocity of the tracked entity sufficiently to get it back to the center of the camera field quickly with minimal overshoot.

The stereo disparity error reported by the electronic tracking module is used to drive the vergence control motor on the active head using a simpler proportional control algorithm to keep the cameras verged on the tracked target.

5.2.3 The figure stabilization submodule

The above tracking mechanism is responsive and tracks moving subjects fairly well. Errors in the optical flow tracking, however, are cumulative and the optical flow tracker is also fooled by objects rotating in place. To mitigate these types of tracking effects without degrading the tracker's responsiveness a simple figure-ground discriminator was developed.[16]

The figure-ground discriminator applies an additional set of 6 stereo disparity correlations covering a disparity range of 24 pixels at each of 16 locations on a 4 by 4 grid centered on the tracked window. This grid initially covers about half of the camera field. At each of the 16 locations checked, the stereo correlations are assessed to decide whether or not there is surface material within a disparity range of plus or minus 12 pixels. If so, it is assumed that the tracked figure extends out beyond that image position.

The center of mass of the figure as estimated on the 4 by 4 grid of samples is used to compute an offset bias that if large enough will be used to readjust the tracked window position.

In addition, the approximate size of the tracked figure is computed from its image on the 4 by 4 sample grid. If the grid spacing is to large compared with the estimated object size, the dispersion of the grid is reduced and vice versa.

5.3 Demonstrations

A sequence of demonstrations was prepared during the course of research to assist with testing and performance evaluation. The first applied the basic tracking mechanisms described in section 5.2 to the problem of following a person walking around in a room; the second set of demonstrations placed the tracking system on a mobile outdoor robot and used it to track a fixed landmark while on the move and to guide the robot in a person following demonstration.

5.3.1 Person tracking

The current implementation of the tracker module achieves fairly good performance. In an office environment, it is able to follow an individual walking at a brisk pace at ranges from 20 feet to about 4 feet and over the active

head's full pan and tilt ranges which are plus or minus 190 degrees in pan and plus or minus 30 degrees in tilt.

At present the tracker module does not have any concept of a person's shape or behavior built into it. It has a simple model of figure, implicit in the figure-ground discriminator, and attempts to follow such figures. Thus when following a person who ducks behind some other object, the tracker locks onto the new object and proceeds to follow it.

In the future we plan to augment the basic tracker module with another control layer which will attempt to represent object persistence. This will enable us to recognize the possible occlusion of a tracked entity and anticipate its reappearance elsewhere.

We also are developing control algorithms that will guide the tracking window onto a persons' head once it has locked onto some part of his body. This mechanism will make use of heuristic knowledge such as the fact that in most instances, heads are at the top. This capability will allow us to integrate the tracking system with a face recognition system.

In addition to finding heads, we are investigating mechanisms for initially directing the attention of the tracker onto objects. Our current idea about this is to scan the visual space looking for for things that have disparate motion or that have changed in range from a prior baseline.

5.3.2 Mobile operation

To experiment with visual tracking while in translational and rotational motion, the Prism-3 processor box and the active stereo head was mounted on an outdoor mobile robot as shown in figure 27. This platform measured approximately 1.2 by 1 meters and was capable of driving at about 3 meters per second on level terrain. The mobile system can operate untethered controlled by radio modem.

The stereo head was mounted on a support that placed it near the front of the vehicle at about one meter off the ground. From that position the cameras were above all other parts of the robot giving it a clear 360 degree view.

Two experiments were carried out with this mobile configuration. In the first the tracker was pointed at a fixed landmark with the goal of maintaining track of that target while the mobile platform was in motion. In the second test, the tracker was pointed at a person and the mobile platform was driven

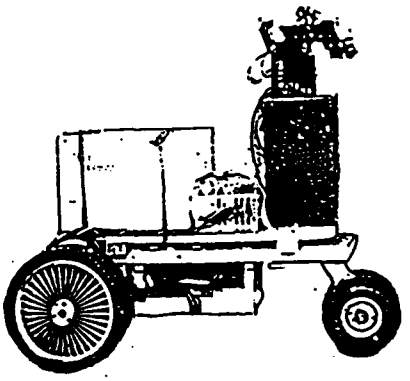

Figure 27: Mobile robot platform with Prism-3 system.

to follow that person guided by the heading and range measurements from the tracker.

Landmark monitoring on the move. A frequent problem in mobile robot navigation is to determine the position of the robot based on known landmarks in the environment. To this end, we attempted to use the tracking system to follow a series of fixed objects (tree trunks, parked cars, etc) and continuously report their position. The results of this experiment proved the robustness of the sign correlation algorithm under a wide variety of demanding outdoor viewing conditions, target textures, and platform motions.

Mobile robot person follower. A natural extension of the indoor person tracking work described in the previous section is to integrate it onto the mobile platform, and then use it to maneuver the platform close to the human target. Such a system is interesting because of the large number of redundant degrees of freedom to be controlled and coordinated to accomplish smooth tracking.

The implementation of the cart control consists of a pure pursuit controller, as described in [17], for turn rate, and a proportional controller for

speed. Essentially, pure pursuit is a proportional controller where turn rate is a function of heading error and distance to the target. Speed is simply a linear function of the target distance, and is clamped at 2.5 m/s for safety. Communication between the vision system and platform controller occurs over ethernet, which adds a variable latency to the issuing of motion commands; typically, this latency is about 10 msec.

Our experience with this system has been very promising. With little tuning, we have been able to achieve smooth and reliable following and approach motion. We have also seen the value in having very agile camera mounts on mobile robot systems. All mobile robots suffer from being relatively slow because of their size and weight, and would be unable to keep up with the dynamic motions of real-world targets. A separate pan-tilt-vergence mechanism, however, mitigates this limitation, simplifying system control and increasing tracking performance.

5.4 Future Work

The tracking systems described here illustrate the power that can be achieved using minimal appropriate measurements in a compact demand directed system. This facility is serving as a base for further research in object and functional recognition making use of the ability to focus sensor attention and maintain track on moving bodies. Results of this work will be applied to task domains that include unmanned vehicle guidance and security and surveillance.

Articulated sensing with fast visual feedback imparted a surprisingly life-like behavior to our tracking systems. This responsive behavior also dramatically increased the performance of the mobile robot system, since the articulated head allowed lower bandwidth control of the mobile robot.

A major thrust of our ongoing research is to build on this reactive sensing capability in the direction of understanding visual persistence. A particularly striking characteristic of the human visual system is the stability it imparts to our perception of the visual world despite ongoing changes in the position of objects and of our head and eyes. This stability, which seems to be achieved by largely autonomous or subconscious mechanisms, helps to make the visual world much easier to deal with. For example, we can look at an object, then glance away so that it is out of our field of view, and then glance back and reacquire its position and description with little effort even when the

object or observer are in motion. In many circumstances we are able to deal with many such objects simultaneously, keeping track of where they are and varying the amount of attention we give each dynamically according to the demands of the task at hand.

6 UGV technology development

During the course of this research program, many tasks were carried out in direct support of the UGV program. These include: taking the lead for investigating the possible role of narrow-field-of-view (NFOV) stereo on the UGV; investigating the use of NFOV stereo to produce high resolution wide-field-of-view range maps; studying the feasibility of using stereo-derived landmarks to support navigation; data collection and evaluation of algorithm performance in UGV relevant environments; and collaboration and technology transfer activities with the system integration contractor and the other stereo contractors on the UGV program.

6.1 Narrow-field-of-view-stereo

One of our primary focuses at Teleos has been on active visual perception and we have taken the lead on the UGV program for exploring ways to apply active visual perception in support of meeting UGV mission requirements. In particular, we have identified the following tasks:

1. Far-look-ahead obstacle detection using stereo. As illustrated in figure 28, a narrow-field-of-view stereo sensor can be configured to look for hazards out toward the horizon along the expected direction of vehicle travel.

2. Active following of navigational features. For example, when driving on a road with a steep embankment to one side, a narrow-field-of-view (NFOV) stereo sensor can be used to monitor the position of that embankment.

3. Double checking wide-field-of-view (WFOV) data. The higher resolution pixel data on a NFOV system can be used to check potential hazards detected by the WFOV system.

6.2 Wide-field high-resolution stereo

The first of the above areas to be investigated carefully has been the building of high resolution stereo range mosaics by scanning an active NFOV sensor

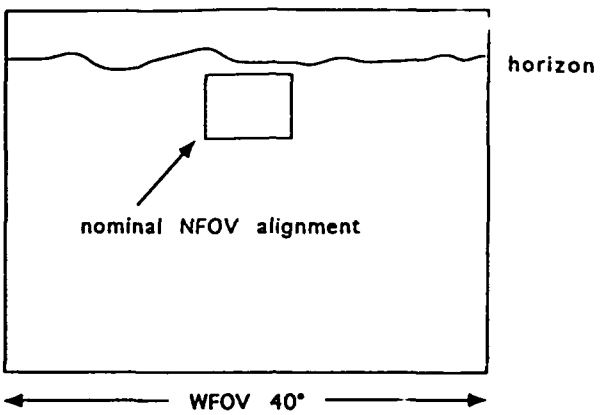

Figure 28: This figure depicts the superposition of wide-field and narrow-field stereo sensors to provide wide coverage and far lookahead in a more restricted zone in the direction of travel.

over a large scene. Figure 29 shows the kind of result that can be obtained. In this figure, the stereo head is looking out a window at an outdoor scene with trees, buildings, and an intersection. The upper image in the figure shows the gray level mosaic image covering about 100 degrees in pan angle and 30 degrees in tilt. The lower images show stereo range using shades of gray, with lighter indicating closer to the cameras. The hatched areas mark locations where the matcher did not find a satisfactory correlation peak. These are largely on the sky which was cloudless. The distances of the close objects (light grey tree crowns) vary from approximately 10 to 30 feet; the distances of the far background (dark grey buildings and trees) range from 70 to 150 feet.

The sample interval in both dimensions is .25 degrees and the spatial resolution of the range measurements is about .5 degrees, or about one foot at one hundred feet. It is interesting to note that objects like the cobra light pole at the upper right are more discernible in the range image than in the r. / camera image.

Figure 29: The upper image shows a mosaic of many camera images, the middle display shows stereo derived range using shades of gray, the bottom display shows the same information with the grayscale adjusted to highlight things farther away.

6.3 Stereo landmarks

As shown in the previous section, an active stereo range sensor equipped with narrow field of view lenses can be used to acquire high-resolution wide field of view range data from the environment around a vehicle. This capability presents the opportunity to track movements relative to its environment, and thus assist local navigation[18]. It may also be feasible to use occupancy maps built from this data to recognize and navigate through previously traversed locales.

Figure 30 shows the effect of viewer motion on the WFOV range data. The top image here is a mosaic gray level image showing a scene covering a pan range of approximately 127 degrees. The middle image is the corresponding range map and the bottom image is a range map computed after moving the sensor 12 feet to the right. This is in the direction of the right edge of the display. The direction of the largest tree in the middle display (top-left of center) is at right angles to the direction of motion. It is half off the image on the left side of the lower display. This data, when analyzed appropriately, can be used to accurately recover the motion the viewer made between images.

We reviewed potential methods for tracking a vehicle's movement through a locale based on this type of range data. We then implemented a voting-based system and obtained promising results when this method was applied to real outdoor stereo data.

The current system could be readily extended to a more general method of navigation. In addition to the localization of the robot within a locale, the present system could be used to identify the current locale of the robot from a database of multiple possible locales. The range map of each distinct locale could be matched to the range map measured at the current position, and the height and sharpness of the peak of each match could be used to determine the correct one. This locale recognition capability could be especially useful in discovering path cross-over points during an extended journey. Maintenance of a database of previously seen locales would also be valuable for guiding a robot back to its point of origin or to any other point along the path it has traversed.

Intelligent feature extraction from the range data, possibly augmented with other sensed information such as color, could potentially increase the performance of the system significantly. We chose, however, to leave this

for later research because the task of defining and actually recovering stable landmark features from range images, like that shown in figure 30, appeared to be beyond the scope of our initial project.

6.4 Test and evaluation

Test and evaluation of stereo algorithms on realistic imagery is an important component of the UGV stereo research program. During the contract period Teleos was involved in developing techniques for recording data and in collecting test data from the Demo A site. Special forms of stero sensor data were also collected and evaluated including night time imagery using intesified cameras and FLIR data. Experiments using a vertical stereo baseline wree also carried out. Finally, several mobile platforms were developed for carrying out live testing of our real-time algorithms.

6.4.1 Recording techniques

A data recording capability sufficient to capture live stereo video on a moving vehicle along with associated navigation and camera attitude data is on the critical path for our research effort. A high performance recording system of this type will take some time to be designed and installed. In the mean time we explored interim methods for recording stereo imagery that might be implemented quickly and inexpensively at all sites so that early test imagery can be collected and shared easily by the UGV stereo contractors.

Several approaches to solving the recording problem were formulated. One involves collecting stereo data by interleaving left and right camera images on the even and odd fields of a single interlaced video frame. This sacrifices half the vertical resolution but should be sufficient to drive early performance studies. To be usable, it will be necessary to acquire the left and right half fields simultaneously while sending them out to the recorder sequentially. Likewise, it will be necessary to undo that encoding on playback for real-time processing experiments. We have implemented this approach on our system.

In assessing the usability of various analogue recording techniques, signal degradation measurements were made by taking test patterns stored in a frame buffer and recording them on video tape. These recorded images were then played back and digitized back into a frame buffer. The resulting digital

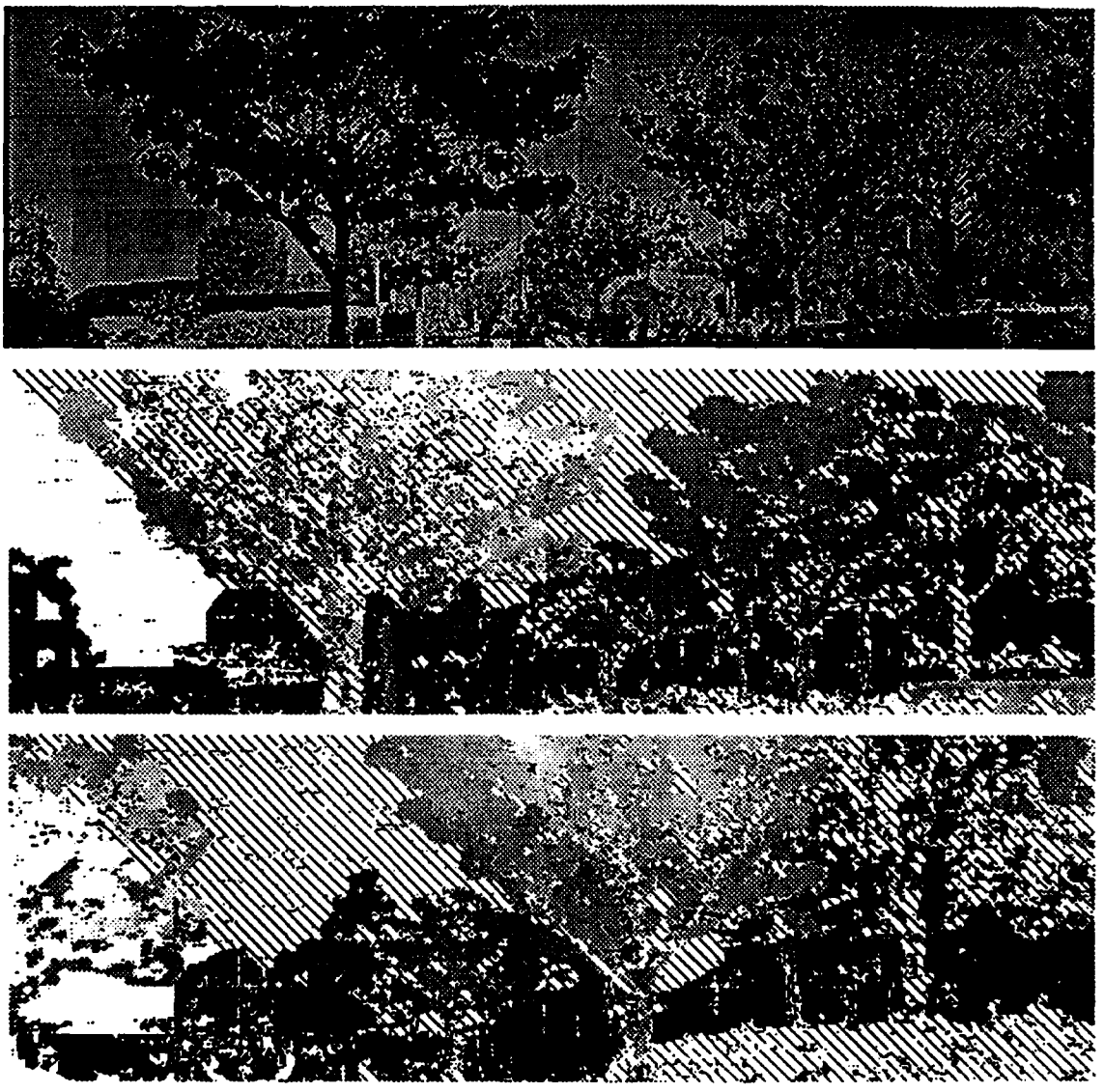

Figure 30: The effect of moving the stereo cameras 12 feet to the right is shown in the change between the middle and bottom range images shown here.

images were then compared with the original test patterns to assess the degree of resolution loss going through video tape storage. These experiments indicated that a loss of a factor of two or three in horizontal resolution is incurred. While this is not an ideal situation, the restored images should be adequate for initial studies working with live video. We also looked at the spatial resolution of and crosstalk between red, green, and blue video channels decoded from VCR tape. As expected, the cross talk was much too large to be of use for recording stereo video.

6.4.2 Demo-A site data

At the 10th UGV workshop at the Martin Marietta facility in Denver, Colorado, SRI and Teleos collaborated to collect stereo video data from the Demo A test course at the Martin Marietta facility. With the assistance of Martin Marietta personnel a HMMWV was outfitted with a roof carrier for stereo cameras, recorders, and a portable generator. The local weather was unstable, shifting from blizzard conditions to balmy blue skys, but a full day's recording was accomplished over both paved road and off-road paths on the test site. The data collected varied a number of parameters including lens focal length, camera base line, vehicle speed, and camera pointing angle.

6.4.3 Low-Light (intensified) stereo

We visited Robotic Systems Technology (RST) in Hampstead, Maryland along with Bob Bolles, SRI, and Larry Matthies, JPL, to collect video data from the stereo-intensified cameras that they have on their Surrogate Teleoperated Vehicle. RST gave us access to their equipment and had one of their staff members assist us. Data was collected in the field after dark using a pickup truck rigged with a portable generator, cameras and recorders. We were able to collect a large set of video test data.

We have done some preliminary checks on the data with our Laplacian-of-Gaussian convolver hardware to see how much stable texture could be picked out through the camera noise. It appears to be a challenging task in total darkness, but we may be able to achieve usable results with very large operators and convolution windows. We made the following observations from the experiments:

1. The intensified cameras produce very large amounts of shot noise in

their outputs. This noise increased dramatically as the scene illumination levels were reduced from lighting from nearby street lights, to lighting from our trucks break lights, to lighting from distant street lights, and finally from star light.

2. Stereo matching using sign-correlation with large convolution and correlation operators could operate to some extent in all but the lowest lighting level.

3. Large gain swings occurred between the two intensified stereo cameras due to their having independent AGC circuits. This was caused by one camera seeing something bright like a distant street light before the other one did. The large gain adjustment this caused would sometimes cause a contrast reversal between the left and right images. It would be desirable in future implementations to couple the AGC circuits to eliminate this kind of problem. Ideally, the camera gains would be computer controlled so that small bright lights could be ignored.

6.4.4 FLIR stereo

A set of static stereo pairs was collected at Martin Marietta using a single Amber sensor mounted on a horizontal slide that allowed taking images from several baseline positions. This data was collected during the day. The data was analyzed using the sign-correlation algorithm and the following observations were made:

1. Texture contrast was quite good on a road scene and these daylight FLIR images were very easy to match in stereo.

2. Night-time data is needed to determine feasibility of FLIR stereo for UGV operation in darkness. This data collection should be done at many different periods, e.g. at dusk, several hours after sunset, before dawn, and just before sunrise. Our intuition is that the most difficult case will come from the before dawn FLIR data since at that time object temperatures in a natural scene will have come closest to equilibrium with the mean air temperature.

3. The observations about coupling or computer controlling stereo camera gains for intensified sensors should apply equally well to the use of FLIR

cameras.

6.4.5 Vertical baseline stereo

Bob Bolles, SRI, suggested at the Fall 93 UGV workshop that a vertical baseline might improve stereo matching performance especially when trying to detect holes in front of the vehicle. Teleos evaluated this proposal using the Prism-3 stereo system and found that matching performance is indeed improved when looking out toward the horizon over a roughly horizontal surface. In this kind of imaging situation, there is often a predominance of horizontal image structure which is useless to a stereo matcher with a horizontal baseline. However, in the case of a vertical stereo baseline, those horizontal edges provide useful matches. In addition, most cameras currently provide significantly more pixel resolution along their horizontal scan directions. This means that a vertical baseline (with the cameras rotated 90°) will increase resolution in the vertical direction which is where it is needed for discrimination of range discontinuities over obstacles.

This result has been reported to Martin Marietta and they have changed their design plans to give the narrow-field-of-view stereo camera head a vertical baseline.

6.4.6 Mobile testbed facilities

To support testing and algorithm development with live video from a moving vehicle, Teleos developed two testbed facilities. A car mounted imaging platform, and a mobile robot platform.

In the first, the Prism-3 real-time system was adapted for operation on a car's roof top to allow simulation of operation on the UGV vehicle close to the Teleos facility. This system has been used for data collection and experimentation and will in the future be used to make live processing runs while driving the vehicle on local back roads with topographical features similar to those at the Demo B site in Denver.

A second mobile platform was developed to allow off-road testing in a smaller space. In this case, a robotic cart was configured to carry the Prism-3 stereo and motion system on outdoor experiments as described in section 5.3.2. This test system allows complete closed loop control of the vehicle and will be used in the future to test the interaction of sensing and control in

high speed obstacle avoidance experiments.

6.5 Collaboration and technology transfer

During the contract period, Teleos has actively particpated in numerous meetings with the integrating contractor, Martin Marietta and our counterparts at SRI and JPL to discuss issues related to the support of stereo sensing on Demo A through Demo II. Teleos also carried out feasibility studies for porting stereo algorithms to parallel processors like the IWARP which had been earmarked for use on the Demo II vehicle. We also provided technical enhancements to the TCX communications interface that was adopted by the integrating contractor for use on the UGV vehicles. Finally, we contributed a proposal for using speech synthesizers to make the operation of stereo sensing during demonstrations more accessible to the audience.

6.5.1 Meetings

In May 1992, Teleos and SRI hosted a UGV stereo review meeting attended by stereo researchers from Teleos, SRI, and JPL, along with contract representative Connie Gray from the U.S. Army Topographic Engineering Center. At the meeting we presented highlights of each group's approaches; we examined some of the stereo video test data collected thus far, including a look at this data through our Laplacian-of-Gaussian convolution hardware; we reviewed some of the results from our (static) stereo evaluation project; and we discussed how best to carry out technology transfer with Martin Marietta. This last discussion led to an e-mail document which was iterated between the three stereo contractors and then sent to Martin Marietta for their comments.

In June 1992, Teleos hosted a meeting with the UGV stereo contractors and Dave Morgenthaler and Dave Anhalt of Martin Marietta to discuss the design of the stereo hardware systems on the first UGV vehicle being assembled by Martin Marietta. SRI, JPL, and Teleos reviewed their ongoing research in stereo and objectives for the program. Discussions ensued about Martin Marietta's proposed stereo hardware architecture vis-a-vis the imaging and processing requirements we anticipate for accomplishing the UGV stereo sensing mission. The meeting was productive and set the stage for a good working relationship between the four groups.

In September 1992, Teleos prepared for and attended the NIST-sponsored Workshop on Performance Evaluation of UGV Technology. The meeting yielded a productive discussion of the issues and challenges of measuring performance in the development of UGV technology.

In October 1992, Teleos hosted David Anhalt from Martin Marietta for a day to review the designs of our stereo algorithm and our accelerator hardware in preparation for the UGV workshop held during October in Denver, CO.

In July 1993 Teleos carried out a live demonstration of the Prism-3 real-time stereo and motion tracking system at the Demo A demonstration at Martin Marietta's Denver facility. This presentation was done collaboratively with the SRI and JPL stereo team members. There was significant interaction with the Demo A attendees and the overall effort was very productive. In preparation for the Demo A presentation, several software enhancements to the to the PRISM3 stereo/motion system were introduced to better demonstrate its capabilities. In addition to work at Teleos, collaborators at NASA's Johnson Space Center contributed a new figure-ground discrimination module that further improved the human tracking performance of the Prism-3 system.

On September 14, 1993 Keith Nishihara attended a meeting at JPL with the UGV stereo contractors and Martin Marietta personnel to discuss the UGV Demo B mission scenario and the role stereo sensing will play. Since there will be no other hazard sensing system available in time for Demo B, stereo will play a critical role in the demonstration. A prioritized list of stereo capabilities was discussed and requirements to support those were formulated in preparation for the UGV workshop in Killeen Texas.

A major area of discussion was about sensor head requirements. The discussion with Martin Marietta is still in progress but the present concept is that there will be a wide field of view (WFOV) pair of color cameras on the navigation pan-tilt mount. A second narrow field of view stereo sensor could be mounted on an independent pan-tilt mount, which could be either the RSTA head or the ALVIN head. It was decided that vergence control on the NFOV stereo cameras could be omitted without serious loss of performance. This change makes the implementation much easier. The possibility of using a vertical baseline on the NFOV cameras was also raised. This could potentially lead to enhanced performance in detecting holes at distance. It also increases the likelihood of the NFOV system being able to recover range

information in regions where the WFOV system is returning low confidence data.

Chuck Shoemaker, ARPA Program Manager, and Paul Lescoe, U.S. Army TACOM, visited Teleos at the beginning of December 1993 for a review of the stereo research program.

In addition, Teleos has actively participated in all of the UGV workshops held during 1992 and 1993.

6.5.2 IWARP port analysis

While the IWARP processor was under consideration for use in the UGV program, we developed test code to assess the feasibility of porting our sign-correlation algorithm to that computation engine. We determined that our pipelined Laplacian-of-Gaussian hardware design could be mapped to the IWARP in a relatively straightforward manner. Timing benchmarks run on the IWARP at CMU indicated that a 64 cell IWARP could do the convolutions at about 1/3 the speed of our video rate convolver board. Analysis of the correlation stage of our algorithm showed that a 64 cell IWARP implementation would run significantly faster than our current single board correlator.

6.5.3 TCX development support

The TCX interprocess communications protocol developed at CMU was adopted for use on the Demo II program. During the contract period, Teleos converted its robotic systems for carrying out UGV research to use TCX. In the process, regular contributions were made to ths shared TXC facility at CMU in the form of bug fixes, enhancements, and ports to new platforms such as the PC and OS9 operating systems.

6.5.4 Self-narrating processes

Teleos developed a proposal for improving user and observer awareness of the internal activity going on in our UGV systems. The proposal was accepted and will be implemented on the UGV Operator Control Unit (OCU) by Hughes for experimentation. The basic idea was to provide speech synthesizers on the various system modules that would provide "self-narration" during the execution of the demonstration. These might be on different audio

channels so that the observer can switch back and forth, or, with inspiration, we might be able to have a number of speakers on a single channel. The messages would generally be canned "printf" statements with enough parameters filled-in to give a running account of what is going on in a given module. The attraction of this idea is that it doesn't take up the operator's visual attention and could lend a fast-paced feel to our demos, which might otherwise appear to be running in slow motion, especially from a distance.

7 Conclusion

This report reviewed the work done from December 1991 to December 1993 at Teleos Research in support of ARPA's UGV program. We have discussed five aspects of our program: task directed visual perception; stereo core research; shape recognition core research; active vision core research; and UGV technology development.

Section two gave a broad overview of Teleos' approach to studying visual perception. In it the concept of minimal-meaningful-measurement tools for early vision was described as a natural methodology for allowing a higher level application process to easily influence and exploit basic measurement modalities. Key to this were the ideas of (1) defining early measurement problems in a minimalist way so that only as much as is necessary to answer basic useful question is computed; (2) structuring the measurement module to have an easy to model behavior so that a user is better able to exploit it in new situations without having to understand the details of the internal algorithm; and (3) providing richer information about that minimal measurement, for example correlation peak shape and height along with the disparity of the peak center. The sign-correlation algorithm under development at Teleos was described in this context and the current performance benchmarks of our accelerator technology were presented.

Section three then discussed our core research program in stereo vision. Highlights of the work during this period include the development of several new algorithms for enhancing stereo matcher performance on UGV relevant imagery. In particular, Teleos developed: (1) techniques for automatically setting stereo matcher operating parameters such as filter size by previewing results on a sparse set of points over a range of possible parameter settings; (2) an analysis identifying the principal parameters affecting the magnitude of the disparity gradient effect that compromises correlator performance in UGV stereo imagery; and (3) a technique for improving area correlator performance in the presence of large stereo disparity gradients.

An equally important component of our stereo core research effort has been in developing methodologies and tests for evaluating the performance of our algorithms in realistic contexts. To this end, Teleos collaborated with the other UGV stereo contractors on a broad evaluation of stereo matching algorithms. A suite of stereo imagery for testing matcher performance in the presence of increasing noise and in the presence of large disparity gradients

was contributed to the project. Teleos also submitted its own stereo matching algorithm to the evaluation process and that algorithm performed well overall and was noteworthy among all compared for its noise handling capabilities.

Section four described progress in representing shape in visual data to support recognition. An overview of Teleos' approach to thinking about shape recognition was presented followed by a report on a technique based on the idea of describing shapes using simple blob primitives. This work led to the realization that circular blob primitives could do the job previously done by a larger family of oriented elliptical shapes and a report is given on research results following that line.

Section five presented results on the application of our work in the real-time domain. An active vision sensor head is described along with a high-performance hardware accelerator for the sign-correlation algorithm. This system is applied to the problem of tracking moving people using stereo and motion sensing modalities and a number of techniques are described for enhancing the speed and robustness of the tracker. This tracking system is then deployed on a mobile outdoor vehicle and used to drive it in a closed loop person following demonstration.

Section six then described activities that related the core research results to specific UGV applications. Studies were carried out in the areas of narrow-field-of-view stereo; wide field-of-view, high-resolution stereo-mosaic building; the use of stereo landmarks in support of vehicle navigation; test and evaluation of stereo algorithms on various kinds of imagery relevant to the UGV mission; and technology transfer support to the UGV system integrator, Martin Marietta, and to other collaborating UGV contractors.

In conclusion, 1992 and 1993 have been productive years for Teleos Research, and much ground work has been laid for continuing work on the UGV program. Future plans include extensive on vehicle experimentation with obstacle detection and avoidance at high speed and extending our core research program to explore the concept of persistent visual objects. Close collaboration with the other UGV stereo contractors will also be continued with an emphasis on integrating our ideas and transferring results to the UGV system integrator.

References

[1] H. K. Nishihara. Minimal meaningful measurement tools. Technical Report TR-91-01, Teleos Research, 1991.

[2] D. Marr and T. Poggio. A computational theory of human stereo vision. *Proceedings of the Royal Society of London*, 204:301–328, 1979.

[3] H. K. Nishihara. Hidden information in transparent stereograms. In *Proc. of the Twenty-First Asilomar Conf. on Signals, Systems, and Computers*, pages 695–700, Pacific Grove, CA, Nov 1987.

[4] H. K. Nishihara. Practical real-time imaging stereo matcher. *Optical Engineering*, 23(5):536–545, October 1984. Also in Readings in Computer Vision: Issues, Problems, Principles, and Paradigms, edited by M.A. Fischler and O. Firschein, Morgan Kaufmann, Los Altos, 1987.

[5] D. Marr and H. K. Nishihara. Representation and recognition of the spatial organisation of 3d shapes. *Proceedings of the Royal Society of London*, 200:269–294, 1978.

[6] H. Blum. Biological shape and visual science, part 1. *J. Theoretical Biology*, 38:205–287, 1973.

[7] J. L. Crowley and A. C. Parker. A representation for shape based on peaks and ridges in the difference of low-pass transform. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-6(2):156–170, 1984.

[8] L. A. Sorayama. Localization of depth edges in stereo images. Master's thesis, Massachusetts Institute of Technology, August 1984.

[9] H. K. Nishihara. Psychophysical and computational tests comparing the sign-correlation and zero-crossing models of human stereo vision. In *Optical Society of America, Image Understanding and Machine Vision, 1989 Technical Digest Series*, volume 14, pages 40–43. North Falmouth, Cape Cod, Massachusetts, June 12-14 1989.

[10] H. K. Nishihara. Tests of a sign correlation model for binocular stereo. *Investigative Ophthalmology and Visual Science*, 30(3):389, 1989.

[11] D. Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.

[12] R. Bajcsy. Active perception. In *Proceedings of the IEEE*, volume 76, pages 996–1005, 1988.

[13] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. In *International Journal of Computer Vision*, number 1, pages 333–356, 1988.

[14] E. Krotkov. *Active Computer Vision by Cooperative Focus and Stereo.* Springer Verlag, New York, 1989.

[15] M. J. Swain and M. Stricker. Promising directions in active vision. Technical Report CS 91-27, University of Chicago, November 1991.

[16] E. Huber. Object tracking with stereo vision. In *AIAA Conference on Intelligent Robots in Field, Factory, Service and Space*, Houston, March 20-24 1994.

[17] R. Wallace, A. Stentz, C. Thorpe, H. Moravec, W. Whittaker, and T. Kanade. First results in robot road-following. In *Proceedings IJCAI*, pages 1089–1095, August 1985.

[18] J. B. Burns and H. K. Nishihara. Feasibility of using stereo to assist navigation. Technical Report TR-92-02, Teleos Research, 576 Middlefield Road, Palo Alto, CA 94301, 1992.